Advances in AI-Generated Images and Videos

Hessen Bougueffa[1], Mamadou Keita[1], Wassim Hamidouche[2], Abdelmalik Taleb-Ahmed[1], Helena Liz-López[3], Alejandro Martín[3], David Camacho[3], Abdenour Hadid[4] *

[1] Laboratory of IEMN, CNRS, Centrale Lille, UMR 8520, Univ. Polytechnique Hauts-de-France (France)
[2] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR, Rennes, 6164 (France)
[3] Computer Systems Department, Universidad Politécnica de Madrid (Spain)
[4] Sorbonne Center for Artificial Intelligence, Sorbonne University Abu Dhabi (United Arab Emirates)

* Corresponding author: bougueffaeutamenehessen@gmail.com (H. Bougueffa), Mamadou.Keita@uphf.fr (M. Keita), whamidouche@gmail.com (W. Hamidouche), abdelmalik.taleb-ahmed@uphf.fr (A. Taleb-Ahmed), helena.liz@upm.es (H. Liz-López), alejandro.martin@upm.es (A. Martín), david.camacho@upm.es (D. Camacho), abdenour.hadid@ieee.org (A. Hadid).

Received 13 November 2024 | Accepted 20 November 2024 | Early Access 28 November 2024

Abstract

In recent years generative AI models and tools have experienced a significant increase, especially techniques to generate synthetic multimedia content, such as images or videos. These methodologies present a wide range of possibilities; however, they can also present several risks that should be taken into account. In this survey we describe in detail different techniques for generating synthetic multimedia content, and we also analyse the most recent techniques for their detection. In order to achieve these objectives, a key aspect is the availability of datasets, so we have also described the main datasets available in the state of the art. Finally, from our analysis we have extracted the main trends for the future, such as transparency and interpretability, the generation of multimodal multimedia content, the robustness of models and the increased use of diffusion models. We find a roadmap of deep challenges, including temporal consistency, computation requirements, generalizability, ethical aspects, and constant adaptation.

Keywords

AI-Generated Content, Image Generation, Multimodal, Video Generation.

DOI: 10.9781/ijimai.2024.11.003

I. Introduction

The recent progress in Artificial intelligence (AI) has led to a revolution in the creation of synthetic images and videos, mainly due to the remarkable capabilities of advanced generative models, diffusion models, or Generative adversarial networks (GANs), among others. There are now a large number of applications and tools available to users, such as DALL-E [1], GLIDE [2], Midjourney [3], Imagen [4], VideoPoet [5], Sora [6], or Genie [7]. These tools are designed to produce realistic and believable digital content easily. This development has had a profound impact, with various applications across different areas.

These techniques are capable of generating multimedia content on any topic or object. Therefore, there are countless opportunities, especially in some application domains, which can benefit greatly from these techniques and tools: *entertainment and media*, allowing the generation of characters, scenarios or elements that would be very difficult to create by traditional means [8]–[10]; *creative industries*, allows artists to streamline their work and improve its quality, for example by creating sketches to work on further, or creating elements to add to their work [11], [12]; *education*, creating engaging educational content, including simulations and visual aids to help illustrate and clarify complex ideas, and adapting to different learning styles [13], [14]; *security and forensics*, helping to create robust models capable of detecting false or generated information more easily, for example by assisting in data augmentation [15], [16]. As we can see, the applications of these techniques are limitless, and as their capabilities improve, they can be more easily applied to different problems in society.

This collection of tools and methodologies not only presents advantages, but also a number of weaknesses and potential risks that need to be carefully analysed. The ability to produce highly realistic synthetic media easily causes concern about their possible inappropriate use. Deepfakes and other kinds of manipulated content can be used to spread misinformation, create disinformation, and manipulate public opinion, undermining trust in digital media [17], [18]. This dual potential for both positive and negative impact highlights a crucial problem. While leveraging the benefits of generative models, there is an urgent need to develop effective detection methods to distinguish between real and AI-generated content. As generative models become more sophisticated, the task of detecting synthetic media becomes increasingly complex, necessitating the continuous evolution of detection techniques.

Please cite this article as:
H. Bougueffa, M. Keita, W. Hamidouche, A. Taleb-Ahmed, H. Liz-López, A. Martín, D. Camacho, A. Hadid. Advances in AI-Generated Images and Videos, International Journal of Interactive Multimedia and Artificial Intelligence, (2024), http://dx.doi.org/10.9781/ijimai.2024.11.003

Despite the significant advancements in generative models, several gaps and challenges persist in both their deployment and the methods used to detect synthetic media. One major challenge lies in the resource-intensive nature of training and deploying these models. High computational requirements limit accessibility, particularly for smaller organizations and researchers lacking the necessary infrastructure to fully utilise these technologies. This creates a barrier to wider adoption and raises concerns about the scalability and sustainability of generative models as they continue to evolve. Furthermore, even advanced models such as GLIDE [2] and DALL·E 2 [1] encounter challenges when processing complex prompts. These challenges can limit their ability to generate high-quality outputs under specific conditions. Similarly, Imagen [19] enhances computational efficiency but still grapples with resource demands and complex prompts. These limitations underscore a need for improved flexibility and robustness in current generative technologies.

On the video generation front, text-to-video models face significant challenges in maintaining high fidelity and continuity of motion over extended sequences. Many existing methods simply extend text-to-image models, which do not fully address the unique complexities inherent in video generation. This highlights the need for more specialized approaches that can effectively handle the temporal dynamics and continuity required for high-quality video content.

Detecting synthetic media presents significant challenges. Current detection models struggle to keep pace with the rapid advancements in generative technologies, making it difficult to reliably differentiate between real and AI-generated images and videos. These models tend to specialize in the types of synthetic content they were trained on, leading to poor performance when faced with new data from different or updated models. Additionally, detection algorithms must be resilient against various transformations and adversarial attacks [20], [21], such as image compression and blurring, which can significantly diminish their effectiveness. Techniques for identifying deepfakes [22] and other forms of image and video forgeries [23] also encounter obstacles due to the constantly evolving nature of these manipulations and the need for high-quality datasets and standardized benchmarks.

To address these challenges and advance the field, this survey:

- Presents an updated picture of synthetic image generation and detection techniques.
- Presents an overview of video generation and detection techniques.
- Provides a list of the main video and image datasets used by researchers.
- Describes trends, challenges and research directions that can be explored in the AI generation, in video and image, and supports them with the conclusions of the analysis.

By providing a thorough examination of both the generative and detection aspects of synthetic media, this survey aims to foster a deeper understanding of the current challenges and opportunities in the field, promoting the development of technologies that can maximize the benefits of AI-generated content while minimizing its risks.

This survey is structured to comprehensively address both the generative capabilities and detection techniques of AI-generated images and videos, see Fig. 1. Section II reviews related works and surveys, providing a foundation for understanding the current state of research in this domain. Section III dives into image generation and detection, detailing various advanced generative models and the methods used to detect synthetic images. Section IV focuses on video generation and detection, exploring the advancements in video generation and the techniques to identify AI-generated videos. Section V discusses the datasets used for generative and detection algorithms, highlighting the importance of diverse and high-quality datasets. Section VI identifies the ongoing challenges in both generating and detecting synthetic media. Finally, Section VII concludes the survey by summarizing the key findings and suggesting future directions for research and development in this field.

II. Related Work and Related Surveys

The field of AI-generated images and videos has been extensively studied, with several surveys reviewing the advancements and challenges in this area. This section provides an overview of key surveys and positions our work in relation to them, highlighting the unique aspects of our approach, summarised in Table I.

- Liu *et al.* [24] conducted an extensive review on human image generation, categorizing existing techniques into three main paradigms: data-driven, knowledge-guided, and hybrid. The survey covers the most representative models and approaches within each paradigm, highlighting their specific advantages and limitations. Additionally, it explores a range of applications, datasets, and evaluation metrics relevant to human image generation. The paper also addresses the challenges and potential future directions in the field, offering valuable insights for researchers interested in this rapidly evolving domain.
- Chen *et al.* [28] concentrated on controllable text-to-image generation models. They investigated various methods that precisely control the produced content, such as personalized and multi-condition generation techniques. The authors explore the practical applications of these models in content creation and design while also recognizing current constraints and suggesting future directions to enhance the adaptability and accuracy of these generative models.
- Joshi *et al.* [29] provided an extensive analysis on the use of synthetic data in human analysis, focusing on the advantages and challenges in biometric recognition, action recognition, and person re-identification. The survey delves into various techniques for generating synthetic data, including deep generative models and 3D rendering tools, emphasizing their potential to tackle issues related to data scarcity, privacy concerns, and demographic biases in training datasets. Additionally, the authors explore how synthetic data can augment real datasets to enhance model performance scalability analysis and simulate complex scenarios that are challenging to capture with real data. They also address concerns about synthetic datasets, such as identity leakage and lack of diversity.
- Figueira *et al.* [25] focused on the generation of synthetic data with Generative Adversarial Networks (GANs). The authors emphasize the significance of synthetic data, particularly in cases where data is limited or contains sensitive information. They highlight how GANs can proficiently create high-quality synthetic samples that imitate real data distributions. This study presents a detailed summary of current methods and challenges in synthetic data generation, emphasizing the utilization of GANs for diverse data types, including tabular data, and exploring various GAN architectures that cater to these requirements.
- Nguyen *et al.* [26] offered a comprehensive review of deepfake generation and detection methods using deep learning techniques. They explored different types of deepfakes, such as face-swaps, lip-syncs, and puppet-master variations, while highlighting the progress and challenges in identifying these manipulations. The survey covers traditional and deep learning-based approaches for detecting deepfakes, including methods based on manual feature creation and those utilizing deep neural networks. Their work emphasizes the importance of developing robust detection algorithms to counter

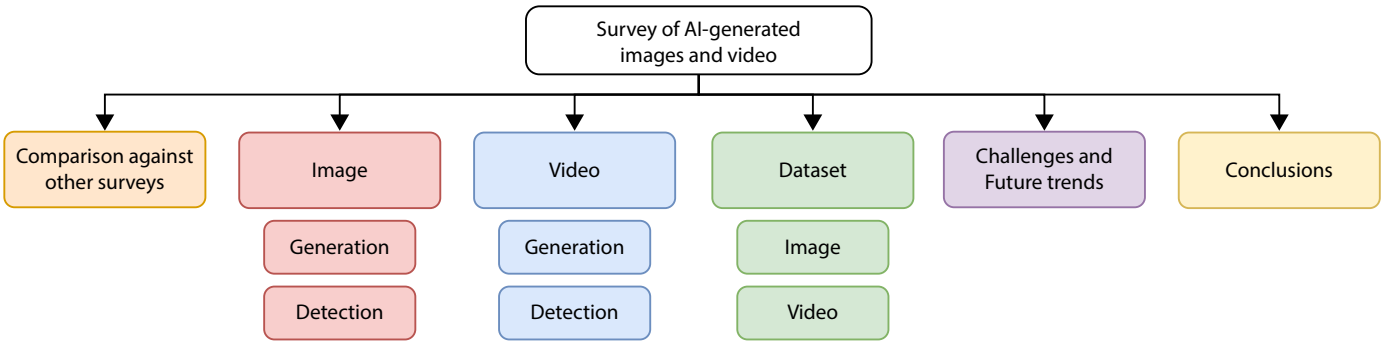


Fig. 1. Schematic representation of the structure followed.

TABLE I. Comparison of Previous Literature Reviews

Authors	Year	Task analysed		Modalities		Main Contribution	Limitations
		Generation	Detection	Image	Video		
Liu *et al.* [24]	2022	✓	×	✓	✓	It provided an extensive review on the generation of human images	It only deals with the generation of human images, without covering other possible scenarios.
Zhang *et al.* [22]	2022	✓	✓			It provides a detailed analysis of video and image sample manipulation and detection techniques.	Focus on the manipulation of video and image samples.
Figueira *et al.* [25]	2022	✓	×	✓	×	It provides a very detailed analysis of the use of GANs within data generation, focusing on training problems and evaluation techniques.	It does not focus on image and video generation.
Nguyen *et al.* [26]	2022	✓	✓	✓	✓	It analyses both the techniques of generation, or manipulation, and the detection of images and videos.	It is mainly focused on the manipulation of multimedia data, not so much on the generation of synthetic samples.
Tyagi *et al.* [23]	2023	✓	✓	✓	✓	Performs a detailed analysis of manipulation and detection techniques for video and audio samples.	The focus is not on synthetic sample generation and detection techniques, but on manipulation techniques.
Bauer *et al.* [27]	2024	✓	×	✓	✓	It performs one of the most comprehensive data generation analyses available.	It is not focused on the generation of image and video samples.
Chen *et al.* [28]	2024	✓	×	✓	×	It covers one of the newest approaches to image generation, diffusion models for Text-to-image task.	This is a very limited survey, as it covers only one of the imaging approaches, without analysing other techniques or modalities.
Joshi *et al.* [29]	2024	✓	×	✓	✓	Explores techniques including improving model performance, increasing data diversity and scalability, and mitigating privacy issues.	It only focuses on generating samples that represent humans, leaving a large part of the field unstudied.

the increasing complexity of deepfake creation techniques. This study holds particular relevance in developing new multimodal approaches for deepfake detection, which are in alignment with investigating cross-modality fusion strategies.

- Bauer *et al.* [27] examined Synthetic Data Generation (SDG) models, analyzing 417 models developed over the past decade. The survey classifies these models into 20 distinct types and 42 subtypes, providing a comprehensive overview of their functions and applications. The authors identified significant model performance and complexity trends, highlighting the prevalence of neural network-based approaches in most domains, except privacy-preserving data generation. The survey also discusses challenges, such as the absence of standardized evaluation metrics and datasets, indicating the need for enhanced comparative frameworks in future research.
- Zhang *et al.* [22] analysed the generation and detection of deepfakes, shedding light on both the progress made and the challenges encountered in this area. They outline two main techniques for creating deepfakes, face swapping and facial reenactment, and discuss the impact of GANs and other deep learning methods. Their work also explores various detection strategies, ranging from biometric and model features to machine learning-based methods. They emphasize the persistent challenges arising from evolving deepfake technologies, the need for high-quality datasets, and the absence of a standardized benchmark for detection methods. This survey is essential for gaining insights into the current state of generating and detecting deepfakes, which present significant challenges to privacy, security, and societal trust.
- Tyagi *et al.* [23] conducted a comprehensive analysis of image and video forgery detection techniques, highlighting the various manipulation methods, such as morphing, splicing, and retouching, and the challenges associated with detecting these alterations in digital media. The survey also reviewed different datasets used for training and evaluating forgery detection algorithms, emphasizing the need for robust, generalized methods capable of detecting multiple types of manipulations across diverse visual datasets. This work provides a detailed examination of both traditional and deep learning-based approaches, illustrating the advancements and limitations in the field of digital media forensics.

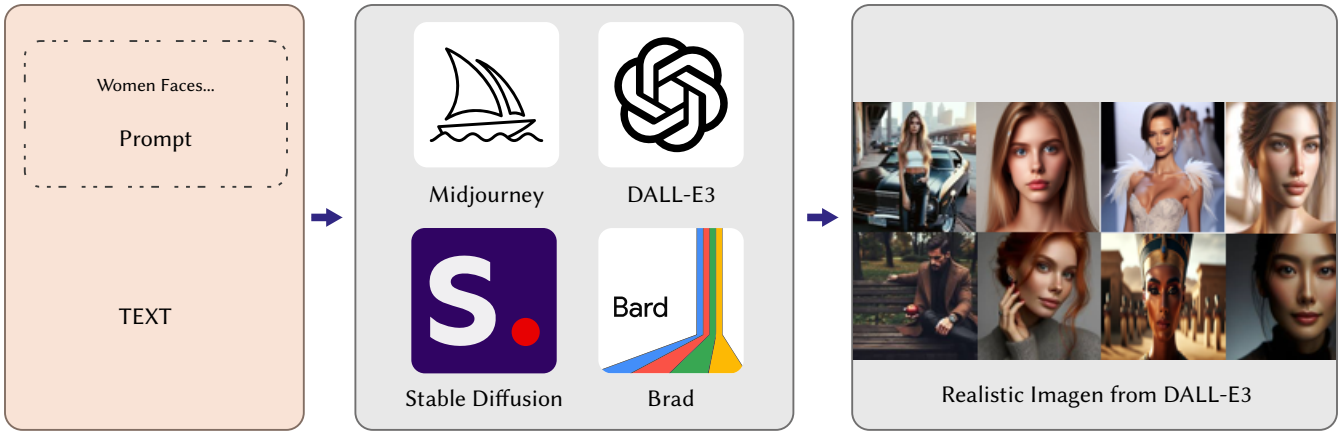


Fig. 2. Overview of the main approaches to image generation with AI.

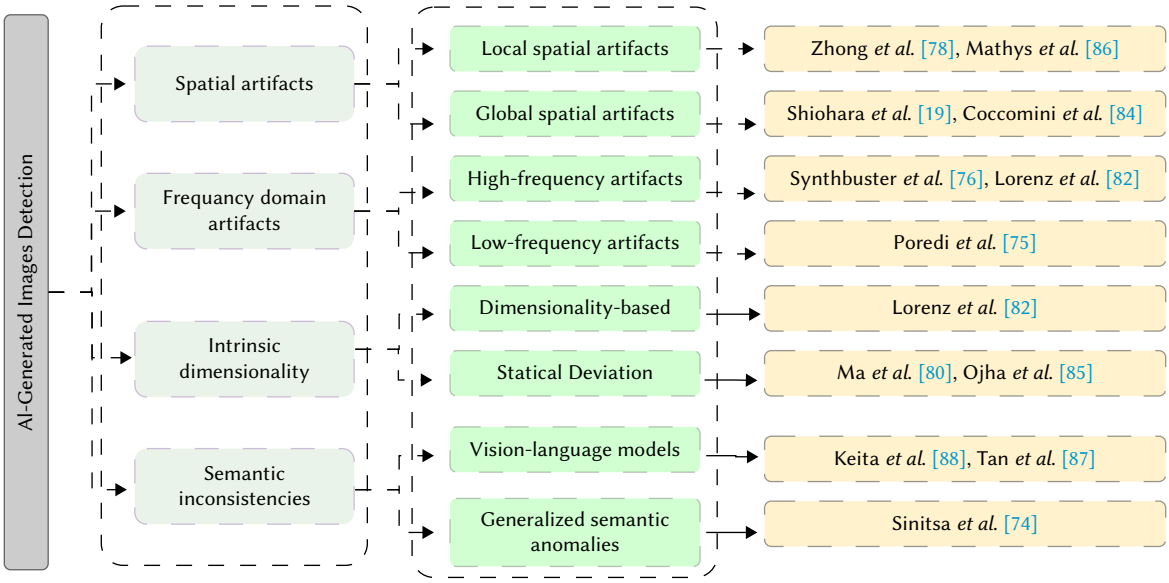


Fig. 3. Overview of AI-generated Image Detection.

As we can see, this survey has a number of advantages over other published reviews of the field. Firstly, it is the first work to focus exclusively on synthetic sample generation techniques, which also provides a list of datasets published in recent years. It also analyses the approaches with which researchers are tackling the problem of detecting these synthetic samples.

III. AI Image Generation and Detection

In this section, we will focus on the generation of images with AI techniques, as well as on the main approaches for their detection. As mentioned above, AI, more specifically Deep Learning (DL) has shown significant progress in the fields of image generation and detection.

Advanced models have greatly improved the ability to **generate synthetic images**, focusing on enhancing aspects such as image quality and realism. Recent developments have led to improved training stability and higher-quality generated images, addressing common challenges and allowing for the creation of diverse and realistic outputs. Innovations in model architectures have also provided greater control over the image generation process, resulting in even more varied and convincing synthetic images. Fig. 2 illustrates a subset of AI-generated image and video techniques, specifically focusing on generative models that rely on text or prompts to create the samples. While this figure highlights key models used in text-to-image or text-to-video synthesis, other generative approaches are discussed in the subsequent sections.

Models for **synthetic images detection** have also made substantial progress. These detection models have become more advanced, using deep learning techniques to identify subtle artifacts and inconsistencies in generated images. As a result, they are crucial in differentiating between real and synthetic images, ensuring the integrity of visual content. The ongoing evolution of these models indicates the dynamic nature of the field, with continuous research efforts focused on improving their precision and resilience [30], [31].

A. Image Generation

Within AI image generation, we will analyse two different approaches, see Fig. 3. The first approach, ***Text-to-image synthesis***, will focus on generating image samples from text descriptions; while the second approach, ***Image-to-image translation***, focuses on modifying the original image while preserving some visual properties in the final sample. A concise summary of the main image generation techniques is presented in Table II.

TABLE II. Comprehensive Overview of a Few Synthetic Image Generation Techniques

Models	Year	Technique	Target Outcome	Data Used	Open Source
NVAE [66]	2020	Hierarchical VAE	High-fidelity images	CelebA, FFHQ	No
CogView [41]	2021	Transformer-based	Text-to-image synthesis	Diverse text and images	Yes
StyleGAN3 [59]	2021	GAN-based	High-quality images	FFHQ, CelebA	Yes
BigGAN [73]	2021	GAN-based	Large-scale image synthesis	ImageNet	Yes
GLIDE [2]	2021	Diffusion-based	Generate images from text prompts	DALL-E's dataset	Yes
DALL-E 2 [1]	2022	Transformer-based	Text-to-image synthesis	Custom, diverse content	Yes
DiVAE [38]	2022	VQ-VAE with diffusion	High-quality reconstruction	ImageNet	No
VQ-VAE-2 [65]	2022	VA E -based	High-resolution images	Large-scale datasets	Yes
EfficientGAN [61]	2022	GAN-based	Efficiency and quality	Custom datasets	Partial
Latent Diffusion [43]	2023	Diffusion-based	Photorealistic images	Various	Yes
DALL-E 3 [51]	2023	Enhanced Transformer	Improved prompt following	Custom image captioner dataset	No
Imagen [4]	2023	Transformer-based	High-fidelity image synthesis	Open Images, ImageNet	No
Imagen2 [50]	2023	Style-conditioned diffusion	Lifelike images with context	Diverse dataset	No
Muse [40]	2023	Transformer T5-XXL	High-fidelity zero-shot editing	CC3M, COCO	No
SDXL [48]	2023	Stable Diffusion	High-resolution image synthesis	Custom dataset	Yes
StyleGAN-T [32]	2023	GAN-based	High-quality image synthesis	Comprehensive dataset with various text-image pairs	Yes
GALIP [35]	2023	GAN-based, utilizing CLIP	Efficient quality image creation from text	Diverse datasets	Yes
GigaGAN [33]	2023	Advanced GAN	High-resolution, detailed image generation from text	Extensive datasets with diverse image-text pairs	Yes
UFOGen [37]	2024	GAN and diffusion	High-quality fast generation	-	No
RAPHAEL [49]	2024	Diffusion with MoEs	Artistic images from text	Subset of LAION-5B	Yes
Ahmed et al. [36]	2024	GAN with spatial co-attention	Enhanced image generation	CUB, Oxford-102, COCO	No

1. Text-to-Image Synthesis

In this section, we will look at different approaches to creating synthetic images from text. As this is a growing field we can observe a variety of different techniques, such as GANs, transformers or diffusion models.

Generative Adversarial Networks: Some authors continue to focus on GANs which, although not particularly novel, have competitive results in the field. For example, Sauer *et al.* [32] have improved the robust StyleGAN architecture to develop StyleGAN-T. This model tackles the challenge of producing visually diverse and attractive images from textual descriptions at scale, effectively speeding up the process while maintaining image fidelity. StyleGAN-T is trained on a comprehensive dataset containing various text-image pairs, ensuring diverse visual outputs. However, one limitation is the potential for reduced accuracy in rendering complex scenes due to the inherent challenges of text ambiguity and the current limitations of GANs in understanding nuanced textual descriptions. Kang *et al.* [33] proposed GigaGAN's, an architecture that includes an improved generator and discriminator that efficiently handle large-scale data, allowing for the creation of diverse and visually compelling images. However, like other large-scale GANs, GigaGAN requires significant computational resources for training and has the potential to overfit precise textual descriptions if the training data lacks diversity. Despite these limitations, GigaGAN's image synthesis capability is a powerful tool in AI-driven creative image generation, expanding the boundaries of machine understanding and visualization of textual content. The model TextControlGAN [34] introduces an innovative method to improve text-to-image synthesis by modifying the Generative Adversarial Network (GAN) architecture. This modification aims to enhance control and precision in generating images from textual descriptions, by integrating specific control mechanisms within the GAN framework. This capability is essential for applications that require high fidelity between textual inputs and visual outputs, such as in digital media creation and automated content generation.

Other authors have explored the option of **a combination between GAN with other types of techniques** such as the CLIP model, such as Ming Tao *et al.* [35] have applied the pre-trained CLIP model to Generative Adversarial Networks (GANs) to transform the process of text-to-image synthesis. This innovative approach enhances the efficiency and quality of the images created from textual descriptions. By integrating CLIP into both the discriminator and generator, the model achieves strong scene understanding and domain generalization using fewer parameters and less training data. By leveraging diverse and extensive datasets, this method enables the generation of a broad range of intricate and visually appealing images. This approach accelerates the synthesis process and ensures a smoother and more controllable latent space, thereby significantly reducing the computational resources typically required for high-quality image synthesis.

Ahmed *et al.* [36] proposed a novel approach that involves simultaneously generating images and their corresponding foreground-background segmentation masks. This is achieved by using a new Generative Adversarial Network (GAN) architecture named COS-GAN, which incorporates a spatial co-attention mechanism to improve the quality of both the images and segmentation masks. The innovative aspect of COS-GAN lies in its ability to handle multiple image outputs and their segmentations from textual descriptions, thereby enhancing applications such as object localization and image editing. It was extensively tested on diverse datasets, including CUB, Oxford-102, and COCO. However, it faces challenges, such as the

high computational demand required for training and potential biases embedded within the large-scale datasets used. These limitations could impact the generalizability and ethical deployment. By contrast, Xu *et al.* [37], chose to combine these GAN with diffusion models. They proposed UFOGen, that offers a novel approach to generating high-quality images from text quickly. Combining elements of Generative Adversarial Networks (GANs) and diffusion models efficiently creates images in a single step, eliminating the need for slower, multi-step processes used by standard diffusion models. UFOGen's training process is greatly improved by utilizing pre-trained diffusion models, which enhances efficiency and reduces training times. However, similar to other generative models, UFOGen also faces limitations. It depends on large-scale datasets that may contain biased or inappropriate content, potentially leading to biased generated images, which raises ethical concerns and affects the fairness and diversity of the output.

Autoencoder models: Another approach we have seen in the generation of images from text is **autoencoder models**. For example, Saharia *et al.* [4] introduced Imagen, a text-to-image model using classifier-free guidance (CFG) and a pre-trained T5-XXL encoder to improve computational efficiency. The model's key innovation is using large language models to enhance image quality and text-image alignment. Imagen generates images starting at 64×64 resolution, then upscales to 256×256 and 1024×1024 using super-resolution models. Despite achieving a strong FID score of 7.27 on COCO, the model faces challenges with dataset biases, high computational demands, and difficulties in generating realistic human images. On the other hand Shi *et al.* [38] developed DiVAE, which combines a VQ-VAE architecture with a denoising diffusion decoder to create highly realistic images, excelling in image reconstruction and text-to-image synthesis tasks. Using a CNN encoder, the model first compresses images into latent embeddings and then reconstructs them into high-quality images through a diffusion-based decoder. Trained on the ImageNet dataset, DiVAE delivers superior performance in terms of FID scores compared to models like VQGAN. However, the diffusion process is computationally intensive, requiring many steps, and the model is restricted by the fixed image size determined by the training data.

Contrastive learning: it has also been shown that this type of learning is a good technique for tackling this type of task using AI models. The CLIP model [39], created by OpenAI, has attracted the attention of a large number of researchers. This model is able to relate images and text by using contrastive learning, training on large multimodal datasets to align visual and linguistic representations in a shared space, allowing tasks such as image generation, search and classification to be performed without the need for specific supervised training. As a result, it is one of the most widely used approaches for researchers to generate synthetic images from text.

Tranformer: We have also analysed different research that has used transformers for the generation of synthetic images. Muse [40] is a Transformer designed for text-to-image generation. It utilizes a pre-trained T5-XXL language model to predict masked image tokens. Trained on 460 million text-image pairs from CC3M and COCO datasets, this model excels in generating high-fidelity images and supports zero-shot editing, such as inpainting and outpainting. Muse's efficiency exceeds that of diffusion and autoregressive models due to its discrete token space and parallel decoding. However, it faces challenges in rendering long phrases, handling high object cardinality, and managing multiple cardinalities in prompts. Ming Ding *et al.* [41] have introduced CogView. This model harnesses a 4-billion-parameter Transformer architecture in combination with a VQ-VAE tokenizer. CogView operates by encoding text into discrete tokens, which the Transformer processes to forecast corresponding visual tokens. These visual tokens are then transformed into high-quality images using the VQ-VAE decoder. CogView underwent training on extensive datasets, incorporating image-text pairs from diverse sources. Despite its remarkable capabilities, CogView does have limitations. The model demands substantial computational resources for training owing to its expansive parameter size. Similar to numerous text-to-image models, it encounters challenges with intricate or ambiguous text prompts, leading to less precise image generation. Additionally, dependence on extensive datasets can introduce biases within the training data, impacting the variety and impartiality of the generated images. CogView2 [42] used a sophisticated Transformer architecture to quickly generate high-quality images from text. The model begins by producing low-resolution images and then progressively refines them using super-resolution modules, ensuring detailed and consistent results. With a foundation built on a 6-billion-parameter Transformer, the model is trained on diverse datasets of text-image pairs, allowing it to handle tasks such as text-to-image generation, image infilling, and captioning in multiple languages. Nevertheless, CogView2 requires substantial computational resources and careful tuning to balance local and global coherence in the generated images.

Diffusion models. This is one of the topics that has attracted the most researchers. Latent Diffusion Models (LDMs) [43] are a major step forward in high-resolution image synthesis, see Fig. 4. They achieve this by using diffusion models within the latent space of pre-trained autoencoders. This reduces the computational requirements typically associated with diffusion models operating in pixel space while maintaining high visual fidelity. Incorporating cross-attention layers within the UNet backbone is a significant advancement in LDMs. It enables the generation of high-quality outputs based on various input conditions, such as text prompts and bounding boxes. This architecture supports high-resolution synthesis using a convolutional approach. The model is trained to predict a less noisy version of the latent variable by focusing on essential semantic features rather than on high-frequency details that are often imperceptible.

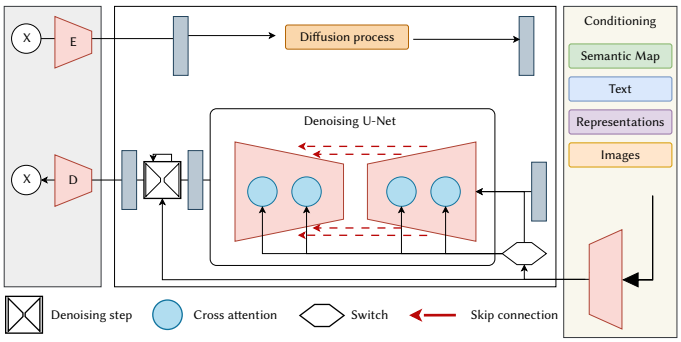


Fig. 4. Latent Diffusion Models architecture from Rombach *et al.* [43].

Anton *et al.* [44] present a new method for synthesizing images from a text by combining image-prior models with latent diffusion techniques. The model utilizes CLIP to map text embeddings to image embeddings and incorporates a modified MoVQ implementation as the image autoencoder. After training on the COCO-30K dataset, Kandinsky achieves high-quality image generation with a competitive FID score. Despite the need for further improvements in the semantic coherence between text and generated images, Kandinsky's versatility in supporting text-to-image generation, image fusion, and inpainting represents a significant advancement in AI-driven image synthesis. EmoGen [45] marks a significant leap forward in text-to-image models. It centers on producing images that capture distinct emotions, solving the difficulty of linking abstract emotions with visual representations. This model excels at creating images that are semantically clear and resonate emotionally. It accomplishes this by aligning the emotion-specific space with the powerful semantic

capabilities of the CLIP model. This alignment is established through a mapping network that interprets abstract emotions into concrete semantics, guaranteeing that the generated images faithfully reflect the intended emotional tones. The model has undergone training and validation using EmoSet, a comprehensive visual emotion dataset with detailed attribute annotations, aiding in optimizing the model for diverse and emotionally accurate image generation. Despite its advancements, EmoGen faces challenges akin to other generative models, including reliance on potentially biased large datasets and the substantial computational resources needed for training and inference, limiting its accessibility and applicability across different research groups and practical uses.

Latent Diffusion Models (LDMs) also have their limitations. One significant challenge is the use of large-scale, often uncurated datasets, which can introduce biases and ethical concerns. While LDMs are more computationally efficient than traditional pixel-based diffusion models, they still require substantial computational resources for training and inference, which may be prohibitive for smaller research groups. LDMs also struggle with generating realistic images of people, leading to lower preference rates in evaluations. Additionally, these models can reflect societal biases, highlighting the importance of robust bias mitigation strategies and the need for more ethically curated datasets in future research. Hang Li et al. [46] present an innovative approach focusing on the ethical implications of AI-generated content and introduce a self-supervised method for identifying interpretable latent directions within diffusion models. The objective is to mitigate the generation of inappropriate or biased images, thus enhancing control over the generated images and ensuring they align with ethical standards while avoiding perpetuating harmful stereotypes. The model has been trained on diverse datasets, allowing it to handle a broad scope of concepts sensitively and responsibly. However, the extensive reliance on datasets may introduce potential biases, while the high computational demand for processing these datasets presents challenges for accessibility and scalability.

Some researchers have chosen to combine the CLIP model with diffusion models. For example, Nichol *et al.* [2] introduced GLIDE, a text-to-image diffusion model that replaces class labels with text prompts. It uses classifier guidance, with a CLIP model in noisy image space, and classifier-free guidance [47], which integrates text features directly into the diffusion process. GLIDE's 3.5B parameter model encodes text through a transformer to generate high-quality images. While effective in photorealism and caption alignment, GLIDE struggles with complex prompts and requires substantial computational power. Ramesh *et al.* [1] introduced DALL·E 2, a model leveraging CLIP and diffusion techniques for generating realistic images from text descriptions. DALL·E 2 operates in two stages: a prior model creates a CLIP image embedding from text, followed by a diffusion-based decoder that generates the final image. This architecture ensures both diversity and realism in the output. The model's use of CLIP embeddings captures semantic and stylistic nuances, enabling high-quality image generation and manipulation. Although trained on a vast dataset, DALL·E 2 faces challenges with complex prompts and fine-grained attribute accuracy, highlighting areas for further improvement.

Furthermore, Podell *et al.* [48] developed SDXL, which is a major step forward in high-resolution image synthesis, expanding on the foundational work of Stable Diffusion models. It utilizes a significantly larger UNet backbone, about three times larger than its predecessors, with more attention blocks and a larger cross-attention context. This enhanced architecture enables SDXL to tackle complex text-to-image synthesis tasks effectively. Additionally, SDXL incorporates multiple innovative conditioning schemes and is trained on various aspect ratios, enhancing its versatility in producing images of different resolutions and aspect ratios. Firstly, it generates initial 128×128 latents. Then, a specialized high-resolution refinement model is applied to improve these latents to higher resolutions. The SDXL training involved utilising an improved autoencoder from previous Stable Diffusion versions. It exceeded its predecessors in all assessed reconstruction metrics, ensuring improved local and high-frequency details in the generated images. The final training stage included multi-aspect training with different aspect ratios, further boosting the model's capabilities. Despite its progress, SDXL has some limitations. The model's reliance on large-scale datasets can lead to biases and ethical concerns due to potentially inappropriate content such as pornographic images, racist language, and harmful social stereotypes. SDXL also struggles to create realistic images of people, often resulting in lower preference rates. Furthermore, the model perpetuates existing social biases, favouring lighter skin tones. Xue *et al.* [49] presents Raphael, an innovative method for generating images from text. It aims to create highly artistic images that closely match complex textual prompts. The model stands out for its mixture-of-experts (MoEs) layers, incorporating both space-MoE and time-MoE layers, allowing for billions of unique diffusion paths. This distinct approach enables each path to function as a "painter," translating individual parts of the text into corresponding image segments with high fidelity. RAPHAEL has outperformed other state-of-the-art models like Stable Diffusion and DALL-E 2. It excels in generating images across diverse styles, such as Japanese comics and cyberpunk, and has achieved impressively low zero-shot FID scores on the COCO dataset. Training on a combination of a subset of LAION-5B and some internal datasets has ensured a broad and diverse range of training images and text for RAPHAEL.

Several tools based on diffusion models have also emerged, such as the following:

- **Imagen2** [50]: this model can generate realistic images by improving the way it pairs images with captions in its training data. The model is adept at understanding context and can edit images, including inpainting and outpainting. It also offers style conditioning, allowing for the use of reference images to guide style adherence, providing greater flexibility and control. However, it struggles with complex object placement and specific detail generation, and there is a possibility of biased content, so safety measures are essential. Trained on a large and diverse dataset, Imagen2 achieves high-quality, contextually aligned image generation.
- **Dall-E3** [51]: has made significant strides in text-to-image generation through the use of improved image captions to enhance prompt following. By developing a custom image captioner to generate detailed, synthetic captions, the model has greatly improved its ability to follow prompts, coherence, and the overall aesthetics of the generated images. However, DALL-E 3 still grapples with issues such as spatial awareness, object placement, unreliable text rendering, and the tendency to hallucinate specific details like plant species or bird types. The model's training consists of a mix of 95% synthetic captions and 5% ground truth captions, which helps regulate inputs and prevent overfitting. This thorough training process allows DALL-E 3 to produce high-quality images with improved prompt following and coherence.

As we have seen in this section, we have analysed the different approaches that are currently being researched within the domain of text-to-image synthesis. The most commonly used techniques have been GANs, Transformers, Diffusion Models and the CLIP model. This shows that there are a large number of synthetic image generation techniques that will allow the creation of large datasets created with many different techniques. This will allow the creation of detection models that are able to generalise better to real situations.

2. Image-to-Image Translation

Recent advances in image-to-image translation have introduced several cutting-edge models that enhance generated images' quality, efficiency, and versatility.

Computer vision is one of the most important fields where GANs are applied, and realistic image generation is the most widely used application of these techniques. For example, Augmented CycleGAN [52] builds on the traditional CycleGAN architecture to handle more complex image-to-image translation tasks, improving domain adaptation, style transfer, and reducing artifacts. DualGAN++ [53] introduces advanced regularization techniques and optimized training strategies, resulting in higher fidelity and fewer distortions in synthetic images. CUT++ [54] refines the original CUT model with contrastive learning techniques and enhanced loss functions for generating higher-quality synthetic images, especially in scenarios with limited data availability. SPADE++ [55] incorporates new strategies for better handling spatial inconsistencies and enhancing the realism of high-resolution synthetic images, particularly effective for images with complex structures. SSIT-GAN [56] leverages self-supervised learning techniques to generate high-quality synthetic images with self-supervised loss functions, useful for applications with limited annotated data. UMGAN [57] proposes a unified approach for multimodal image-to-image translation, enabling the generation of diverse synthetic images from multiple input modalities across various applications. Zero-shot GANs [58] aim to generate images without extensive labelled data, enhancing the zero-shot learning capabilities of GANs. This approach allows for the creation of diverse and high-quality images even with minimal training data.

Recent advancements in GAN-based synthetic image generation have been focused on enhancing image quality, efficiency, and usability across different domains. StyleGAN3 tackles the issue of "texture sticking" in generated images by introducing architectural revisions to eliminate aliasing, ensuring that image details move naturally with depicted objects. The new design interprets all signals continuously, achieving full equivariance to translation and rotation at subpixel scales. This results in images that maintain the high quality of StyleGAN2 but with improved internal representations, making StyleGAN3 more suitable for video and animation generation. The model was trained using high-quality datasets such as FFHQ, METFACES, AFHQ, and a newly collected BEACHES dataset. However, the architecture assumes specific characteristics of the training data, which can lead to challenges when these assumptions are not met, such as with aliased or low-quality images. Additionally, further improvements might be possible by making the discriminator equivariant and finding ways to reintroduce noise inputs without compromising equivariance [59], [60]. EfficientGAN [61] focuses on optimizing computational efficiency while maintaining high-quality image generation. This model aims to reduce the resource requirements for training GANs without compromising the generated images' visual quality. It introduces novel architectural modifications and training strategies that balance performance and efficiency.

Other authors have explored how to combine GANs with other types of techniques such as Latent Diffusion Models, which combine GANs with diffusion models to achieve high-resolution image synthesis. The integration of latent diffusion models helps in generating detailed and high-quality images while maintaining the robustness of GANs [62]. In contrast, Torbunov *et al.* [63] chose to combine them with Transformers. They introduced UVCGAN, an advanced model designed for image-to-image translation, focusing on synthetic image generation. This model improves upon the traditional CycleGAN framework by integrating a Vision Transformer (ViT) into the generator, enhancing its ability to learn non-local patterns. UVCGAN is highly effective for unpaired image-to-image translation tasks, making it a valuable tool for applications in fields such as art, design, and scientific simulations. ViT enables more complex and nuanced image transformations, pushing the boundaries of synthetic image generation possibilities.

Recently, significant developments have been made in Variational Autoencoders (VAEs) for synthetic image generation. These advancements have resulted in the creation of innovative models that enhance the quality, efficiency, and versatility of the images generated. For instance, Conditional VAEs [64] have improved inpainting results and training efficiency by utilizing pre-trained weights and datasets such as CIFAR-10, ImageNet, and FFHQ. VQ-VAE-2 employs hierarchical latent representations to capture high-resolution details, leading to a notable improvement in image fidelity and diversity [65]. NVAE [66], with its hierarchical architecture and advanced regularization techniques, has enabled high-resolution, realistic image generation. Another example is StyleVAE [67], which integrates VAEs with style transfer techniques to produce visually appealing images with stylistic consistency. Additionally, FHVAE has enhanced the disentanglement of latent factors, allowing for better control over image attributes [68]. EndoVAE [69], developed by Diamantis et al., introduces a fresh approach for producing synthetic endoscopic images using a Variational Autoencoder (VAE). This novel technique addresses the drawbacks of traditional GAN-based models, particularly in the domain of medical imaging where maintaining data privacy and diversity is crucial. EndoVAE is specifically designed to generate a diverse set of high-quality synthetic images, which can be used in lieu of real endoscopic images. This aids in the training of machine learning models for medical diagnosis. The outcomes illustrate that EndoVAE adeptly creates realistic endoscopic images, positioning it as a promising tool for advancing medical image analysis and circumventing the challenges stemming from limited data availability.

Furthermore, Dos Santos *et al.* [70] have introduced a Synthetic Data Generation System (SDGS) that utilizes Variational Autoencoders (VAEs) to produce synthetic images. Their system aims to automate the creation of synthetic datasets by using the Linked Data (LD) paradigm to collect and merge data from multiple repositories. The SDGS framework incorporates advanced feature engineering methods to enhance the quality of the dataset before training the VAE model. This results in synthetic images that closely mimic real-world data, making them extremely useful for training machine learning models, especially in scenarios where actual data is scarce. The system's efficacy has been confirmed through various case studies, demonstrating that the generated synthetic data achieves high accuracy and closely resembles the original datasets in crucial characteristics. Seunghwan *et al.* [71] have introduced a new method for creating synthetic data using Variational Autoencoders (VAEs). Their approach overcomes the limitations of the typical Gaussian assumption in VAEs by incorporating an infinite mixture of asymmetric Laplace distributions in the decoder. This advancement provides more flexibility in capturing the underlying data distribution, which is crucial for generating high-quality synthetic data. Their model, known as "DistVAE," has demonstrated exceptional performance in generating synthetic datasets that maintain statistical similarity to the original data and also ensures privacy preservation. The effectiveness of the approach was confirmed through experiments on various real-world tabular datasets, indicating that DistVAE can generate accurate synthetic data while allowing for adjustable privacy levels through a tunable parameter. This makes it particularly valuable in situations where data privacy is a concern.

Finally, we can see how the use of diffusion models in image-to-image translation is also beginning to be explored. For example, Parmar *et al.* [72] proposed pix2pix-zero, a method for image-to-image

TABLE III. Overview of Techniques for Detecting AI-Generated Images

Authors	Year	Technique	Target Outcome	Data Used	Open Source
Shiohara *et al.* [19]	2022	Self-blended images	Detect fake or synthetic images	Self-blended image data	Yes
Wang *et al.* [79]	2023	Diffusion Reconstruction Error	Detect difusión model-generated images	DiffusionForensics dataset	Yes
Ma *et al.* [80]	2023	Deterministic reverse and denoising computation errors	Detect images from difusión models	CIFAR-IO, TinyImageNet, CelebA	Yes
Zhong *et al.* [78]	2023	Texture patch analysis	Identify Al-generated images	Datasets from 17 generative models	Yes
lorenz *et al.* [82]	2023	Intrinsic Dimensionality-based	Detect artificial images from deep diffusion models	CiFake, ArtiFact, DiffusionDB, LAION-5B, SAC	Yes
Alzantot *et al.* [77]	2023	Wavelet-packet representation analysis	Differentiate real and synthetic images	FFHQ, CelebA, LSUN, Face Forensics++	Yes
Poredi *et al.* [75]	2023	Frequency analysis	Identify Al-generated images on social media	Stanford image dataset	Yes
Bammey *et al.* [76]	2023	Frequency artifacts analysis	Detect images generated by diffusion models	Raise and Dresden datasets	Yes
Guarnera *et al.* [83]	2023	Hierarchical classification	Identify deepfake images	CelebA, FFHQ, ImageNet	Yes
Ojha *et al.* [85]	2023	Universal fake image detector	Enhance detection of synthetic or fake images	Images generated by various models	Yes
Mathys *et al.* [86]	2024	CNN-based pixel-level analysis	Identify synthetic images	Diverse dataset with real and synthetic images	No
Coccomini *et al.* [84]	2024	Visual and textual feature classification	Detect synthetic images from diffusion models	MSCOCO and Wikimedia datasets	Yes
Tan *et al.* [87]	2024	Category Common Prompt in CLIP	Enhance detection of deepfakes	Images generated by various models	Yes
Sinitsa *et al.* [74]	2024	Fingerprint-based	Detect synthetic images with low-budget models	Various models datasets	Yes
Keita *et al.* [88]	2024	Vision-language model with dual LORA mechanism	Detect synthetic images using vision-language model	Various datasets	Yes

translation without relying on text prompts or additional training. This approach utilizes cross-attention guidance to maintain image structure and automatically discovers editing directions in the text embedding space. The architecture leverages pre-trained Stable Diffusion models for tasks like object type changes and style transformations. The model's performance is assessed using real and synthetic images from the LAION 5B dataset. However, some limitations include the low resolution of the cross-attention map for fine details and challenges with atypical poses and fine-grained edits.

In this section we have analysed the latest work in the field of Image-to-Image translation, focusing on image alterations while maintaining some visual features. Within this domain we have looked at three main approaches: GANs, AutoEncoders and diffusion models. We can observe that this domain although it has been widely explored, still presents a wide range of possibilities.

B. Detection of AI-Generated Images

The development of generative models requires the creation of detection models to differentiate between AI-generated and real images. Detection methods can be split into two main types: those focused solely on improving detection performance and those that enhance detectors with additional features such as generalizability, robustness, and interpretability while maintaining accurate and effective detection capabilities. An overview of techniques for detecting AI-generated images is provided in Table III, summarizing various methods and their key features, including the application areas and datasets used. For example, the Deep Image Fingerprint (DIF) [74] method is specifically designed to detect low-budget synthetic images. It can identify images generated by both Generative Adversarial Networks (GANs) and Latent Text-to-Image Models (LTIMs). The method utilizes datasets from various models, including CycleGAN, ProGAN, BigGAN, StyleGAN, Stable Diffusion, DALL·E-2, and GLIDE, and achieves high detection accuracy with minimal training samples. While it excels in detecting synthetic images, it may encounter some challenges with models like GLIDE and DALL·E-2 due to their weaker, less distinct fingerprints.

Some authors still opt for more traditional techniques, such as the **Fourier Transform** for the detection of artefacts left in the image samples. For example, the AUSOME (AUthenticating SOcial MEdia) [75] method is focused on identifying AI-generated images on social media. It achieves this by utilizing frequency analysis techniques, such as the Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT), to compare the spectral features of AI-generated images, like those produced by DALL-E 2, with legitimate images from the Stanford image dataset. AUSOME can distinguish between AI-generated and real images by examining differences in frequency responses. Although it demonstrates high accuracy, it may encounter difficulties when dealing with images where semantic content is essential for determining authenticity. Nevertheless, this method presents a promising approach for verifying social media images, particularly in light of the increasing prevalence of AI-generated content. Synthbuster [76] is a technique developed to identify images created by diffusion models by analyzing frequency artifacts in the Fourier transform of residual images. This method is effective at spotting synthetic images, even when they are slightly compressed in JPEG format, and it works well with unknown models. It analyzes real images from the RAISE and Dresden datasets and synthetic images from various models such as Stable Diffusion, Midjourney, Adobe Firefly, DALL·E 2, and DALL·E 3. While Synthbuster is generally effective, it may encounter challenges when dealing with different compression levels and diverse image categories.

Other authors focus on taking advantage of **textures**, in order to exploit all available information. For instance, Alzantot *et al.* [77] proposed multi-scale wavelet-packet representations. Their deepfake image analysis and detection technique aims to differentiate real from synthetic images by analyzing their spatial and frequency information. This method has undergone evaluation using various datasets, including FFHQ, CelebA, LSUN, and FaceForensics++. It has shown strong capabilities in identifying GAN-generated images, such as those created by StyleGAN. However, it may face challenges when analyzing complex images where semantic information is crucial, and its effectiveness may be limited to the detection of image-based synthetic media. PatchCraft [78] introduces a fresh approach to identifying synthetic AI-generated images. Instead of relying solely on global semantic information, this method focuses on analyzing texture patches within the images for more effective detection. To enhance detection, the method employs a preprocessing step called Smash&Reconstruction, which removes global semantic details and amplifies texture patches, thereby utilizing the contrast between rich and poor texture regions to boost performance. Tested on datasets from 17 common generative models, including ProGAN, StyleGAN, BigGAN, CycleGAN, ADM, Glide, and Stable Diffusion, the method has shown superior adaptability and resilience against previously unseen models and image distortions. Nevertheless, it may encounter challenges when dealing with images in which semantic information is critical for accurate detection.

An analysis on the **error** inserted in generated images has also been a productive research line. For example, the DIRE (DIffusion REconstruction Error) [79] method is utilized to identify images created through diffusion processes by comparing the reconstruction error between an original image and its reconstructed version using a pre-trained diffusion model. This technique is based on the idea that diffusion-generated images can be accurately reconstructed using diffusion models, unlike genuine images. DIRE has been evaluated using the DiffusionForensics dataset, encompassing images from various diffusion models, including ADM, DDPM, and iDDPM. It has demonstrated notable accuracy in detecting images and is resilient to unseen diffusion models and alterations. Nonetheless, it may encounter difficulties with the intricate features of real images. Shiohara *et al.* [19] has introduced an innovative approach for detecting fake or synthetic images, specifically deepfakes. They utilize self-blended images (SBIs) as synthetic training data to enhance the robustness of detection models. This allows the models to effectively identify various types of deepfake manipulations by scrutinizing inconsistencies and artifacts in the images. Consequently, this method provides a robust tool for preserving the authenticity of digital media in the face of increasingly advanced generative techniques. The SeDID [80] method utilizes deterministic reverse and denoising computation errors found in diffusion models. This approach includes two branches: the statistical-based SeDIDStat and the neural network-based SeDIDNNs. SeDID was evaluated on various datasets like CIFAR-10, TinyImageNet, and CelebA and demonstrated superior detection accuracy and robustness against unseen diffusion models and perturbations. However, the method may encounter challenges when dealing with the complex features of real images. Nevertheless, SeDID underscores the importance of selecting the optimal timestep to enhance detection performance.

As expected, another approach widely used by state-of-the-art researchers is **Convolutional Neural Networks**, which have demonstrated excellent performance on numerous similar classification problems [81], making it one of the most explored techniques. Some authors continue to rely on classical architectures such as ResNet. It continues to perform competitively on many classification problems. Among them, The multi-local Intrinsic Dimensionality (multiLID) [82] method is developed to identify artificial images produced by deep diffusion models. This method utilizes the local intrinsic dimensionality of feature maps extracted by an untrained ResNet18, making it efficient and not relying on pre-trained models. It has been evaluated on various datasets like CiFake, ArtiFact, DiffusionDB, LAION-5B, and SAC, demonstrating high accuracy in detecting artificial images from models including Glide, DDPM, Latent Diffusion, Palette, and Stable Diffusion. However, multiLID may have limitations in its ability to perform well on unfamiliar data from different datasets or models within the same domain. Guarnera *et al.* [83] developed a hierarchical multi-level approach for detection and identification of deepfake images produced by GANs and Diffusion Models (DMs). This method utilizes ResNet-34 models at three levels of classification: distinguishing genuine images from AI-generated ones, discerning between GANs and DMs, and identifying specific AI architectures. Their dataset comprises authentic images from CelebA, FFHQ, and ImageNet, as well as synthetic images from nine GAN models (e.g., AttGAN, CycleGAN, ProGAN, StyleGAN, StyleGAN2) and four diffusion models (e.g., DALL-E 2, GLIDE, Latent Diffusion), totalling 42,500 synthetic and 40,500 real images. With an accuracy of over 97%, the method demonstrates strong performance, but it may encounter challenges related to real-world robustness, such as JPEG compression and complex image features.

However, other authors have opted for different architectures rather than CNNs. Coccomini *et al.* [84] investigate the detection of synthetic images generated by diffusion models, such as those created with Stable Diffusion and GLIDE. Their approach involves using classifiers like multi-layer perceptrons (MLPs) and convolutional neural networks (CNNs) to distinguish synthetic images from real ones. The model is trained on datasets like MSCOCO and Wikimedia, focusing on leveraging visual and textual features for effective detection. A notable limitation of the study is the challenge of cross-method generalization, where models trained on one type of synthetic image struggle to detect images generated by different methods. This work underscores the complexities of detecting AI-generated images, particularly as diffusion models become more sophisticated. Ojha et al. [85] have introduced a method to enhance the detection of synthetic or fake images generated by various models, including **GANs** and diffusion models. Their approach aims to create a universal fake image detector that performs well across different generative models. This is achieved through a combination of convolutional neural networks (CNNs) and advanced training techniques to identify subtle anomalies commonly found in AI-generated images. The model is trained on diverse datasets, incorporating images generated by various models to improve its reliability. However, the study highlights a challenge in maintaining high detection accuracy when faced with new generative models not included in the training set, indicating the need for further improvements to achieve universal detection capabilities. Mathys *et al.* [86] present a method for identifying synthetic images produced by AI models. The focus is on spotting subtle artifacts and inconsistencies that are indicative of AI-generated content. Their proposed architecture utilizes a convolutional neural network to scrutinize pixel-level details and capture the distinct markers left by generative models. Training the model on a diverse dataset containing both real and synthetic images from various sources makes it adept at generalizing across different types of AI-generated content. This method significantly boosts the accuracy of detecting fake images, effectively tackling the challenges brought about by the increasingly lifelike outputs of modern generative models. This research holds particular significance in upholding the authenticity and integrity of digital content in an age where synthetic media is increasingly prevalent.

Lastly, we will analyse some research that has chosen other novel approaches such as the use of models like **CLIP** or **vision-language**

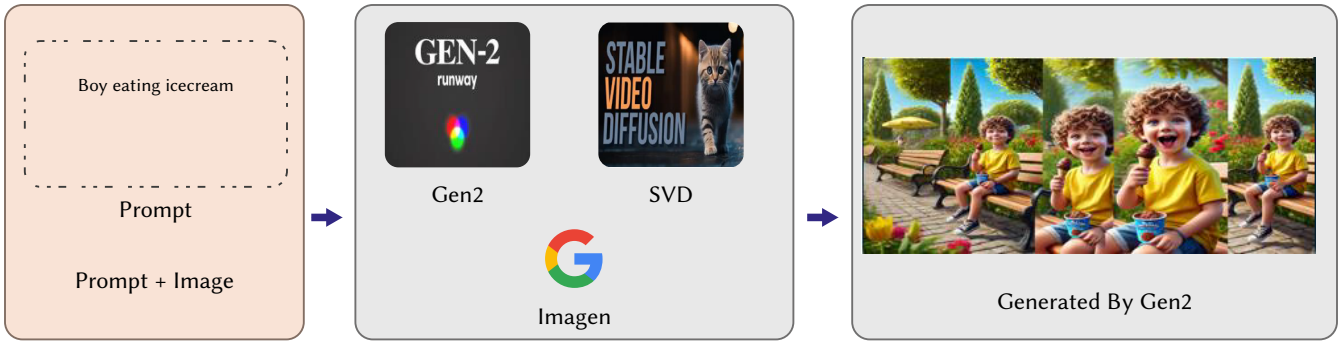


Fig. 5. Overviewof the main approaches to video generation with AI.

models. Tan *et al.* [87] introduce C2P-CLIP, a novel approach designed to enhance the detection of AI-generated images, specifically deepfakes, by injecting a Category Common Prompt (C2P) into the CLIP model. CLIP (Contrastive Language-Image Pre-training) is a powerful model trained on various image-text pairs, which allows it to understand and match images and text descriptions effectively. However, its application to deepfake detection has been limited by its generalization capability across different types of manipulations. The C2P-CLIP method addresses this limitation by incorporating a category-specific prompt that captures standard features across related deepfakes, improving the model's ability to generalize beyond the specific types of manipulations seen during training. This technique leverages the extensive pre-training of CLIP while fine-tuning its capacity to identify subtle inconsistencies and artifacts introduced by deepfake generation techniques. Through comprehensive experiments, the authors demonstrate that C2P-CLIP significantly outperforms existing methods on several benchmark datasets, showing superior performance in detecting a wide range of AI-generated manipulations. Keita et al. [88] present Bi-LORA, a vision-language approach designed to detect synthetic images. Bi-LORA effectively captures the unique features and artefacts of AI-generated images by leveraging a dual Low-Rank Adaptation (LORA) mechanism within a vision-language model. The method integrates visual and textual information, enhancing its ability to differentiate between real and synthetic content more accurately. Through extensive experiments, Bi-LORA demonstrates significant improvements in detection performance over traditional methods, highlighting its potential as a robust tool for identifying AI-generated images across various datasets.

Lastly, we have analysed the most recent research into the detection of synthetic image. This field is highly dependent on the previous one, as quality datasets will be needed, i.e. with intra-class variability, enough quality and resolution, and representativeness, allowing the creation of models that can be used in real situations. In this domain we have seen that the main approaches explored by researchers are CNNs, and vision-language models, although other more traditional approaches are still used.

IV. Video Generation and Detection

In recent years, the field of video generation has attracted significant attention, due to advancements in artificial intelligence, machine learning, and the emergence of diffusion models (see Fig. 5), this has forced researchers to develop new techniques to detect these synthetic samples. This section provides an overview of the current state of video generation methods, which are increasingly being used to create high-quality, realistic videos across different applications. Additionally, it explores the challenges and methods associated with detecting AI-generated videos, an area of growing importance as these technologies become more sophisticated. The aim of this section is to provide a comprehensive understanding of the methods and techniques involved in future video content creation and analysis.

A. Video Generation

In video content creation, generative models are beginning to revolutionize production and consumption by automating the generation of realistic and high-quality videos. Recently, a surge of generative video models capable of various video creation tasks has emerged. In this section we are going to analyse five different approaches: *Text-to-video*, deep learning techniques that generate synthetic video samples from text descriptions; *image-to-video* techniques that transform static images to dynamic video; *video-to-video*, a set of techniques focused on the generation of realistic video sequences by transforming or translating visual information from one video domain to another; *Text-Image-to-Video* which generates synthetic video samples from a real image and a text description; *Multimodal video generation*, this field focuses not only on the generation of the visual part of the video but also on the audio part of the video, from different inputs, such as text, image, video or audio. Deep learning-based generative models such as GANs, Variational Autoencoders (VAEs), autoregressive, and diffusion-based models have remarkably succeeded in generating realistic and diverse content. By training on large datasets, these models learn the underlying data distribution, enabling them to generate samples that closely resemble the original data. Fig. 6 illustrates the various categories of video generation.

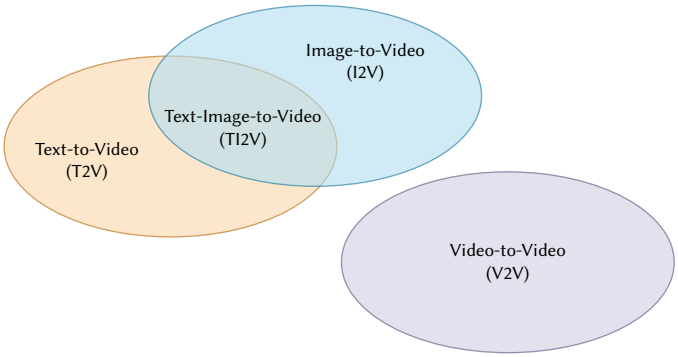


Fig. 6. Categories of video generation methods.

1. Text-to-Video Synthesis

Generating photo-realistic videos presents significant challenges, particularly when it comes to maintaining high fidelity and continuity of motion over extended sequences. Despite these difficulties, recent advancements have utilized diffusion models to enhance the realism of video generation. Text, being a highly intuitive and informative form of instruction, has become a central tool in guiding video synthesis, leading to the development of **Text-to-video** (**T2V**) generation

models. This approach focuses on creating high-quality videos based on text descriptions, acting as a conditional input for the video generation process.

To address the challenges in text-to-video synthesis, existing methods primarily extend **Text-to-image** models by incorporating temporal modules, such as **temporal convolutions and temporal attention**, to establish temporal correlations between video frames. A notable example is the work by Ho *et al.* [89], who introduced Video Diffusion Models (VDM). This model extends text-to-image diffusion models to video generation by training jointly on both image and video data. Their approach utilizes a U-Net-based architecture, which integrates joint image-video denoising losses, ensuring temporal coherence by conditioning on both past and future frames, thus resulting in smoother transitions and more consistent motion. Building on this foundation, Ho et al. [90] proposed Imagen Video, a novel approach for generating high-definition videos using diffusion models. Imagen Video employs a cascaded video diffusion model approach, adapting techniques from text-to-image generation, such as a frozen T5 text encoder and classifier-free guidance, to the video domain. It uses a hierarchical approach, beginning with a low-resolution video to capture the overall structure and motion, which is then progressively refined to higher resolutions. Temporal dynamics are managed by conditioning each frame on previous frames, ensuring consistency throughout the video. Super-resolution techniques are subsequently applied to enhance the detail and quality of each frame.

In a different approach, Singer *et al.* [91] introduced Make-A-Video, which generates videos from textual descriptions without relying on paired text-video data. This methodology builds upon a text-to-image synthesis model and incorporates spatio-temporal layers to extend it into the video domain. The approach integrates pseudo-3D convolutional and attention layers to manage spatial and temporal dimensions efficiently. Additionally, super-resolution networks are employed to improve visual quality, and a frame interpolation network is used to increase the frame rate and smooth out the video output. Meanwhile, Zhou *et al.* [92] presented MagicVideo, a framework designed to generate high-quality video clips from textual descriptions. Instead of directly modeling the video in visual space, MagicVideo leverages a pre-trained **Variational autoencoder (VAE)** to map video clips into a low-dimensional latent space, where the distribution of videos' latent codes is learned via a diffusion model. This approach optimizes computational efficiency and improves video synthesis by performing the diffusion process in the latent space. Further pushing the boundaries of video generation, Dan Kondratyuk *et al.* [5] proposed VideoPoet, an advanced language model for zero-shot video generation. This model integrates the MAGVIT-v2 [93] tokenizer for images and videos and the SoundStream [94] tokenizer for audio, enabling the processing and generation of multimedia content within a unified framework. VideoPoet employs a prefix language model with a decoder-only architecture as its backbone, facilitating the creation of high-quality videos from textual prompts, along with interactive editing capabilities. VideoPoet is trained on a diverse set of tasks without needing paired video-text data, allowing it to learn effectively from video-only examples. It can generate videos based on textual descriptions, animate static images, apply styles [95] to videos through optical flow and depth prediction, and even extend video sequences by iteratively predicting subsequent frames.

In another innovative approach, Girdhar *et al.* [96] introduced EMU VIDEO, a **two-stages Text-to-video generation model**: first, it generates an image from text, and then it produces a video using both the text and the generated image. This method simplifies video prediction by leveraging a pretrained text-to-image model and freezing spatial layers while adding new temporal layers for video generation. EMU VIDEO efficiently achieves high-resolution video generation, maintaining the conceptual and stylistic diversity learned from large image-text datasets. Similarly, Wang *et al.* [97] proposed LaVie, a cascaded framework for Video Latent Diffusion Models (V-LDMs) conditioned on text descriptions. LaVie is composed of three networks: a base T2V model for generating short, low-resolution key frames, a Temporal interpolation (TI) model for increasing the frame rate and enriching temporal details, and a Video super-resolution model (VSR) for enhancing the visual quality and spatial resolution of the videos. The base T2V model modifies the original 2D UNet to handle spatio-temporal distributions and utilizes joint fine-tuning with both image and video data to prevent catastrophic forgetting, resulting in significant video quality improvements. The TI model uses a diffusion UNet to synthesize new frames, enhancing video smoothness and coherence, while the VSR model adapts a pre-trained image upscaler with additional temporal layers, enabling efficient training and high-quality video generation.

Further developments include the work by Menapace *et al.* [98], who proposed a method to generate high-resolution videos by modifying the **Efficient diffusion model (EDM)** [99] framework for high-dimensional inputs and developing a scalable transformer architecture inspired by Far-reaching interleaved transformerss (FITs) [100]. They adjust the EDM framework to handle high SNR in videos with a scaling factor for optimal denoising. This method addresses the scarcity of captioned video data by jointly training the model on both images and videos, allowing for more effective learning of temporal dynamics. The video generation uses FITs, transformer models that reduce complexity by compressing inputs with learnable latent tokens and employing cross-attention and self-attention to focus on spatial and temporal information. The approach includes conditioning tokens for text and metadata and uses a cascade model: the first stage generates low-resolution videos, and the second stage refines them into high-resolution outputs. During training, variable noise levels are introduced to the second-stage inputs to improve upsampling quality, aiming for effective high-quality video generation. In addressing data scarcity, Chen *et al.* [101] designed VideoCrafter2, a model that improves spatio-temporal consistency in video diffusion models through a data-level disentanglement strategy. This approach separates motion aspects from appearance features, leveraging low-quality videos for motion learning and high-quality images for appearance learning. This design strategy eases a targeted fine-tuning process with high quality images, with the aim of significantly increasing the visual fidelity of the generated content without compromising the precision of motion dynamics. Importantly, synthetic images with complex concepts are used for finetuning, rather than real images, to enhance the concept composition ability of video models.

Furthermore, Ma *et al.* [102] introduced Latte, a simple and general video diffusion method that extends **Latent diffusion models (LDMs)** for video generation by employing a series of transformer blocks to process latent space representations of video data obtained from a pre-trained variational autoencoder. Latte specifically addresses the inherent disparities between spatial and temporal information in videos by decomposing these dimensions, allowing for more efficient processing. The method includes four efficient Transformer-based model variants, designed to manage the large number of tokens extracted from input videos, thereby improving the overall performance and scalability of video generation. Li *et al.* [103] introduced VideoGen, a text-to-video generation method that produces high-definition videos with strong frame fidelity and temporal consistency using reference-guided latent diffusion. In their approach, an off-the-shelf T2I model like Stable diffusion (SD) generates a high-quality image from a text prompt, which then serves as a reference for video generation. This process involves a cascaded latent diffusion

module conditioned on both the reference image and text prompt, followed by a flow-based temporal upsampling step that enhances temporal resolution. Finally, a video decoder maps the latent video representations into high-definition videos, improving visual fidelity and reducing artifacts while focusing on learning video dynamics. The training process benefits from high-quality unlabeled video data, using the first frame of a ground-truth video as the reference image to enhance motion smoothness and realism.

Building on the VQ-VAE architecture, Godiva *et al.* [104] proposed GODIVA, an open-domain text-to-video model pre-trained on the HowTo100M [105] dataset. This model generates videos in an auto-regressive manner using a three-dimensional sparse attention mechanism. Initially, a VQ-VAE auto-encoder represents continuous video pixels as discrete video tokens. Subsequently, the three-dimensional sparse attention model utilizes language input alongside these discrete video tokens to generate videos, effectively considering temporal, column, and row information. Similarly, Ding *et al.* [106] advanced the field by introducing CogVideo, a 9B-parameter transformer built upon the pretrained text-to-image model CogView2 [42] for video generation. CogVideo employs a multi-frame-rate hierarchical training strategy, which aligns text with video clips by controlling frame generation intensity and ensuring accurate alignment between text and video content. This is achieved by prepending text prompts with frame rate descriptions, which significantly enhances generation accuracy, particularly for complex semantic movements. Additionally, CogVideo's dual-channel attention mechanism improves the coherence of generated videos by focusing on both textual and visual cues simultaneously. This approach allows CogVideo to efficiently adapt a pretrained model for video synthesis without the need for costly full retraining.

Expanding on the capabilities of earlier models, Wu *et al.* [107] developed NUWA, a unified **multimodal pre-trained model** designed for generating and manipulating visual data, including images and videos, across various visual synthesis tasks. NUWA utilizes a 3D transformer **encoder-decoder** framework to process 1D text, 2D images, and 3D videos. This model introduces a 3D nearby attention (3DNA) mechanism that efficiently handles visual data, reduces computational complexity, and enables high-quality synthesis with notable zero-shot capabilities. Further advancing this work, Wu *et al.* [108] introduced NUWA-Infinity, a groundbreaking model for infinite visual synthesis capable of generating high-resolution images or long-duration videos of arbitrary size. The model features an autoregressive over autoregressive generation mechanism, with a global patch-level model managing inter-patch dependencies and a local token-level model handling intra-patch dependencies. To optimize efficiency, NUWA-Infinity incorpores a Nearby context pool (NCP) to reuse previously generated patches, minimizing computational costs while maintaining robust dependency modeling. Additionally, an Arbitrary direction controller (ADC) enhances flexibility by determining optimal generation orders and learning position embeddings tailored for diverse synthesis tasks. NUWA-Infinity thus transcends the limitations of fixed-size approaches, enabling comprehensive and efficient content creation on a variable scale. In contrast to these approaches, Yan *et al.* [109] proposed VideoGPT, a simpler and more efficient architecture for scaling likelihood-based generative modeling to natural videos. By employing VQ-VAE with 3D convolutions and axial self-attention, VideoGPT learns downsampled discrete latent representations of raw videos. These representations are then autoregressively modeled by a GPT-like architecture with spatio-temporal position encodings to generate videos. This method involves training a VQ-VAE with an encoder that downsamples space-time and a decoder that upsamples it, sharing spatio-temporal embeddings across attention layers. Furthermore, a prior over the VQ-VAE latent codes is learned using an Image-GPT-like architecture with dropout for regularization, which enables conditional sample generation via cross attention and conditional norms. Blattmann *et al.* [110] introduced a novel approach to efficient high-resolution video generation through Video LDMs, by adapting pre-trained image diffusion models into video generators. They achieve this by temporal fine-tuning with alignment layers, which maintains computational efficiency. Initially, an LDM is pre-trained on images and then transformed into a video generator by adding a temporal dimension and fine-tuning on video sequences. Additionally, diffusion model upsamplers are temporally aligned for consistent video super resolution, allowing the efficient training of high-resolution, long-term consistent video generation models using pre-trained image LDMs with added temporal alignment.

Building on these advancements, Chen *et al.* [111] introduced two diffusion models for high-quality video generation: T2V and Image-to-video (I2V). The T2V model, based on SD 2.1, incorporates temporal attention layers to ensure temporal consistency and employs a joint image and video training strategy. The VideoCrafter T2V model further leverages a Latent Video Diffusion Model (LVDM) with a video VAE and a video latent diffusion model, where the VAE reduces sample dimensions to improve efficiency. Video data is encoded into a compressed latent representation, processed through a diffusion model with noise added at each timestep, before being decoded by the VAE to generate the final video. He *et al.* [112] expanded on the concept of video generation by introducing a hierarchical LVDM framework that extends videos beyond the training length. Their method addresses performance degradation with conditional latent perturbation and unconditional guidance. Their lightweight video diffusion models use a low-dimensional 3D latent space, significantly outperforming pixel-space models with limited computational resources. By compressing videos into latents using a video autoencoder and utilizing a unified video diffusion model for both unconditional and conditional generation, their approach generates videos autoregressively and improves coherence and quality over extended lengths with hierarchical diffusion.

To further advance video generation, Wang *et al.* [113] proposed **ModelScope Text-to-Video (ModelScopeT2V)**, a simple yet effective baseline for video generation. This model introduces two key technical contributions: a spatio-temporal block to model temporal dependencies in text-to-video generation, and a multi-frame training strategy with both image-text and video-text paired datasets to enhance semantic richness. ModelScopeT2V evolves from a text-to-image model (stable diffusion) and includes spatio-temporal blocks to ensure consistent frame generation and smooth transitions, adapting to varying frame numbers during training and inference. In the realm of scalable and efficient video generation, Gupta *et al.* [114] proposed W.A.L.T, a simple yet scalable and efficient transformer-based framework for latent video diffusion models. Their approach consists of two stages: an autoencoder compresses images and videos into a lower-dimensional latent space, allowing for efficient joint training on combined datasets. Subsequently, the transformer employs window-restricted self-attention layers that alternate between spatial and spatio-temporal attention, reducing computational demands and supporting joint image-video processing. This method facilitates high-resolution, temporally consistent video generation from textual descriptions, offering an innovative approach to T2V synthesis. Villegas et al. [115] contributed to the field by proposing Phenaki, a unique C-ViViT encoder-decoder structure for generating variable-length videos from textual inputs. This model compresses video data into compact tokens, allowing for the production of coherent and detailed videos. By utilizing a bidirectional masked transformer to translate text tokens into video tokens, the model can generate long, temporally coherent videos from both open-domain and sequential prompts. It

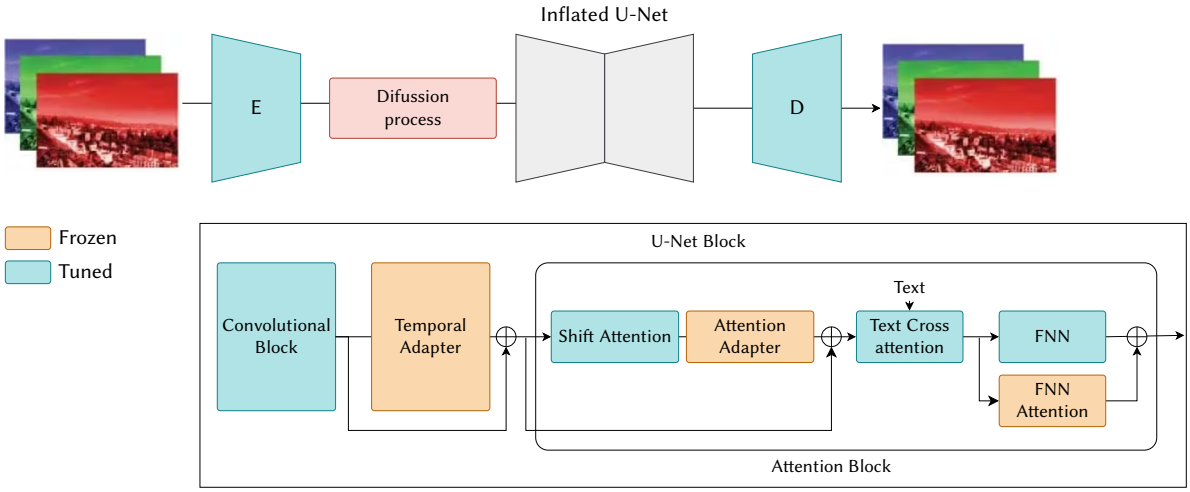


Fig. 7. SimDA [116] architecture.

also improves video token compression by 40% by exploiting temporal redundancy, enhancing reconstruction quality and accommodating variable video lengths, while the causal variation of ViViT manages temporal and spatial dimensions in an auto-regressive manner.

Previous methods of text-to-video generation face high computational costs with pixel-based VDMs or struggle with text-video alignment with latent-based VDMs. To marry the strength and alleviate the weakness of pixel-based and latent-based VDMs, Zhang *et al.* [117] proposed Show-1, a hybrid model that combines both pixel-based and latent-based VDMs to overcome the limitations of previous methods. By employing pixel-based VDMs to create low-resolution videos with strong text-video correlation, and then using latent-based VDMs to upsample these to high resolution, Show-1 ensures precise text-video alignment, natural motion, and high visual quality with reduced computational cost. Khachatryan *et al.* [118] built upon the Stable diffusion T2I model to develop Text2Video-Zero, a zero-shot T2V synthesis model. This approach enriches latent codes with motion dynamics to ensure temporal consistency and employs a cross-frame attention mechanism to maintain object appearance and identity across frames. Although Text2Video-Zero enables high-quality, temporally consistent video generation from textual descriptions without additional training, leveraging existing pre-trained T2I models, there is still potential for improvement. It struggles to generate longer videos with sequences of actions.

Furthermore, FuWeng *et al.* [119] introduced ART•V, an efficient framework for **autoregressive video generation** using diffusion models. ART•V generates frames sequentially, conditioned on previous frames, by focusing on simple, continuous motions between adjacent frames, which helps to avoid the complexity of long-range motion modeling. This approach retains the high-fidelity generation capabilities of pre-trained image diffusion models with minimal modifications and can produce long videos from diverse prompts, such as text and images. To address the common issue of drifting in autoregressive models, ART•V incorporates a masked diffusion model that draws information from reference images rather than relying solely on network predictions, thereby reducing inconsistencies. By conditioning on the initial frame, ART•V enhances global coherence, which is particularly useful for generating long videos. The framework also employs a T2I-Adapter for conditional generation, ensuring high fidelity with minimal changes to the pre-trained model, matching the inference speed of one-shot models, and supporting larger batch sizes during training. In summary, ART•V effectively reduces drifting issues in video generation by incorporating masked diffusion, anchored conditioning, and noise augmentation to better align training with testing. Shi *et al.* [120] introduced BIVDiff, a training-free video synthesis framework that integrates frame-wise video generation, mixed inversion, and temporal smoothing. This framework bridges the gap between specific image diffusion models (e.g., ControlNet, Instruct Pix2Pix) and general text-to-video diffusion models (e.g., VidRD, ZeroScope). The process begins with frame generation using an image diffusion model, followed by Mixed Inversion to adjust latent distributions, which balances temporal consistency with the open-generation capability of video diffusion models. Finally, video diffusion models are applied for temporal smoothing. This method effectively addresses issues of temporal consistency and task generalization that are common in previous training-free approaches.

Finally, Xing *et al.* [116] proposed a parameter-efficient video diffusion model called Simple Diffusion Adapter (SimDA), see Fig. 7, which fine-tunes the large T2I model (i.e., Stable Diffusion) for enhanced video generation. SimDA generates videos from textual prompts through efficient one-shot fine-tuning of pre-trained Stable Diffusion models, focusing on a parameter-efficient approach by fine-tuning only 24 million out of the 1.1 billion parameters. The model employs an adapter with two learnable fully connected layers, incorporating spatial adapters to capture appearance transferability and temporal adapters to model temporal information, utilizing GELU activations and depth-wise 3D convolutions. Additionally, SimDA introduces Latent-shift attention (LSA) to replace the original spatial attention, enhancing temporal consistency without adding new parameters. More recently, Qing *et al.* [121] presented HiGen, a diffusion-based model that improves video generation by decoupling spatial and temporal factors at both the structure and content levels. At the structural level, HiGen splits the T2V task into spatial reasoning, which involves generating spatially coherent priors from text, and temporal reasoning, which creates temporally coherent motions from these priors using a unified denoiser. On the content side, HiGen extracts cues for motion and appearance changes from input videos to guide training, thereby enhancing temporal stability and allowing for flexible content variations. Despite its strengths, HiGen faces challenges in generating detailed objects and accurately modeling complex actions due to computational and data quality limitations.

As we have seen in this section, for the generation of video from text, the main approaches used are the application of T2I techniques together with temporal modules, attention mechanisms, transformers and autoencoder. However, in recent years many researchers are focusing on diffusion models, which are becoming more and more widely used and are expected to increase in popularity in the coming years.

2. Image-to-Video Synthesis

Generating videos from static images poses significant challenges, particularly in preserving temporal consistency and achieving realistic motion across frames. Despite these difficulties, advancements in image-to-video synthesis have leveraged sophisticated modeling techniques to transform still images into dynamic video sequences. This area has become increasingly important for various applications, ranging from content creation to enhanced video editing tools.

Recent methods in image-to-video synthesis focus on generating high-quality videos by incorporating temporal dynamics into the transformation process. Techniques like temporal modeling and attention mechanisms are employed to ensure smooth transitions between frames, thus maintaining coherence and realism in the generated videos. A noteworthy contribution to this field is the work by Wu *et al.* [122], which introduces LAMP, a few-shot-based tuning framework for Text-to-video generation, leveraging a first-frame-attention mechanism to transfer information from the initial frame to subsequent ones. This approach, which focuses on fixed motion patterns, is constrained in its ability to generalize across diverse scenarios. LAMP utilizes an off-the-shelf text-to-image model for content generation while emphasizing motion learning through expanded pre-trained 2D convolution layers and modified attention blocks for temporal-spatial motion learning. A first-frame-conditioned pipeline ensures high video quality by retaining the initial frame's content and applying noise to subsequent frames during training. During inference, high-quality first frames generated by SD-XL enhance video performance. Despite its promise, LAMP faces challenges with complex motions and background stability, suggesting areas for future improvement. Guo *et al.* [123] introduced the I2V-Adapter, a lightweight and plug-and-play solution designed for text-guided Image-to-video generation. The key innovation of this adapter lies in its cross-frame attention mechanism, which preserves the identity of the input image by propagating the unnoised image to subsequent noised frames. This approach ensures compatibility with pretrained Text-to-video models, maintaining their weights unchanged while seamlessly integrating the adapter. By introducing minimal trainable parameters, the I2V-Adapter not only reduces training costs but also ensures smooth compatibility with community-driven models and tools. Moreover, the authors incorporated a Frame Similarity Prior, which provides adjustable control coefficients to balance motion amplitude and video stability, thereby enhancing both the controllability and diversity of the generated videos.

Futhermore, Zhang *et al.* [124] proposed MoonShot, a video generation model that leverages both image and text as conditional inputs. MoonShot addresses limitations in controlling visual appearance and geometry by employing the Multimodal video block (MVB) as its core component. This module integrates spatial-temporal layers for comprehensive video feature representation and utilizes a decoupled cross-attention layer to condition both image and text inputs effectively. Notably, MoonShot reuses pre-trained weights from text-to-image models, allowing for the integration of pre-trained image ControlNet modules to achieve geometry control without necessitating additional training. The model's architecture, which includes spatial-temporal U-Net layers and decoupled multimodal cross-attention layers, ensures high-quality frame generation and temporal consistency. As a result, MoonShot is versatile, supporting tasks like image animation and video editing without the need for fine-tuning, while also enabling geometry-controlled generation through the effective integration of ControlNet modules. Gong *et al.* [125] proposed AtomoVideo, a high-fidelity Image-to-video generation framework that transforms product images into engaging promotional videos. AtomoVideo achieves superior motion intensity and consistency compared to existing methods and can also perform Text-to-video generation by combining advanced text-to-image models. The approach involves using a pre-trained T2I model with added temporal convolution and attention modules, training only the temporal layers, and injecting image information at two positions: low-level details via VAE encoding and high-level semantics via CLIP image encoding and cross-attention. Long video frames are predicted iteratively, using initial frames to generate subsequent ones. The framework is trained using Stable Diffusion 1.5 and a 15M internal dataset, employing zero terminal SNR and v-prediction techniques for stability. During inference, classifier-free guidance with image and text prompts significantly enhances the stability of the generated output.

Other researchers have explored diffusion models for the creation of videos from images. For example, Shi *et al.* [126] proposed Motion-I2V, a novel framework for consistent and controllable text-guided image-to-video generation. Unlike previous methods, Motion-I2V factorizes the process into two stages with explicit motion modeling. The first stage involves a diffusion-based motion field predictor to deduce pixel trajectories of the reference image. The second stage introduces motion-augmented temporal attention to enhance the limited 1-D temporal attention in video latent diffusion models, effectively propagating reference image features to synthesized frames guided by predicted trajectories. By training a sparse trajectory ControlNet for the first stage, Motion-I2V enables precise control over motion trajectories and regions, also supporting zero-shot Video-to-video translation. Although Motion-I2V provides fine-grained control of I2V generation through sparse trajectory guidance, region-specific animation and zero-shot Video-to-video translation, it is limited in handling occlusions, brightness uniformity and complex motion.

Expanding on the idea of temporal consistency, Ren *et al.* [127] proposed ConsistI2V, a diffusion-based method for I2V generation, designed to enhance visual consistency by using spatiotemporal attention over the first frame to maintain spatial and motion coherence. They introduced FrameInit, an inference-time noise initialization strategy that uses the low-frequency band from the first frame to stabilize video generation, which supports applications such as long video generation and camera motion control. The approach leverages cross-frame attention mechanisms and local window temporal layers to achieve fine-grained spatial conditioning and temporal smoothness. The ConsistI2V's architecture, based on a U-Net structure adapted with temporal layers, employs a latent diffusion model to generate videos that closely align with the first frame and follow the textual description. To address motion consistency and efficiency, Shen *et al.* [128] proposed a novel approach to Conditional image-to-video (cI2V) generation by disentangling RGB pixels into spatial content and temporal motions. Using a 3D-UNet diffusion model, they predict temporal motions, including motion vectors and residuals, to improve consistency and efficiency. The approach begins with Decouple-Based Video Generation (D-VDM) to predict differences between consecutive frames and is further refined with Efficient Decouple-Based Video Generation (ED-VDM), which separates content and temporal information using motion vectors and residuals extracted via CodeC. The model employs Gaussian noise and a diffusion model to learn the video distribution score and generate a video clip from the initial frame and text condition. The approach includes Decoupled Video Diffusion Model using DDPM to estimate video distribution scores and a ResNet bottleneck module to encode the first frame, improving spatial and temporal representation alignment. Efficient representation is achieved using I-frames and P-frames, with compression via a Latent Diffusion autoencoder, optimizing video generation through a learned joint distribution of motion vectors and residuals.

Facing the challenge of maintaining temporal coherence while preserving detailed information about the characters in the image-video synthesis for character animation is difficult. Hu *et al.* [129] proposed

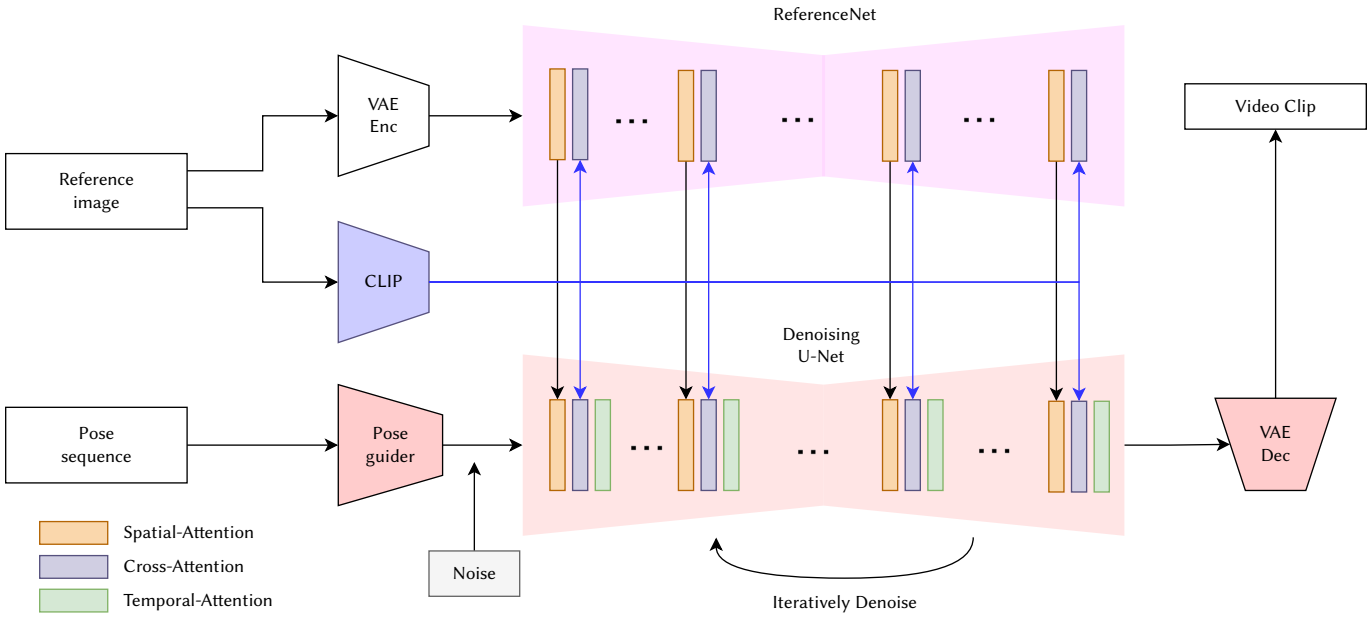


Fig. 8. Animate Anyone [129] architecture.

a novel framework using diffusion models for character animation, see Fig. 8, addressing the challenges of maintaining temporal consistency with detailed character information in image-to-video synthesis. They designed ReferenceNet to merge intricate appearance features from a reference image via spatial attention, and introduced a Pose Guider to ensure controllability and continuity in character movements, along with an effective temporal modeling approach for smooth inter-frame transitions. The method extends Stable Diffusion (SD) by reducing computational complexity through latent space modeling and includes an autoencoder. The network architecture includes ReferenceNet for appearance feature extraction, Pose Guider for motion control, and a temporal layer for continuity of motion. The training strategy consists of two stages: first, training on individual video frames without the temporal layer, and second, introducing and training the temporal layer using a 24-frame video clip. Despite its advancements, the model faces limitations in generating stable hand movements, handling unseen parts during character movement, and operational efficiency due to DDPM. Moreover, Xu *et al.* [130] proposed MagicAnimate, a novel diffusion-based human image animation framework that integrates temporal consistency modeling, precise appearance encoding, and temporal video fusion to synthesize temporally consistent human animation of arbitrary length. They address the challenges of existing methods, which struggle with maintaining temporal consistency and preserving reference identity, by developing a video diffusion model that encodes temporal information with temporal attention blocks and an innovative appearance encoder that retains intricate details of the reference image. MagicAnimate employs a simple video fusion technique to ensure smooth transitions in long animations by averaging overlapping frames. The framework processes animations segment-by-segment to manage memory constraints while leveraging a sliding window method to improve transition smoothness and consistency across segments. This comprehensive approach enables MagicAnimate to produce high-fidelity, temporally consistent animations that faithfully preserve the appearance of the reference image throughout the entire video.

In cases where no motion clue is provided, videos are generated stochastically, constrained solely by the spatial information in the input image. Dorkenwald *et al.* [131] proposed an approach to I2V synthesis by framing it as an invertible domain transfer problem implemented through a Conditional invertible neural network (cINN). To bridge the domain gap between images and videos, they introduced a probabilistic residual representation, ensuring that only complementary information to the initial image is captured. The method allows sampling and synthesizing novel future video progressions from the same start frame. They utilized a separate conditional variational encoder-decoder to compute a compact video representation, facilitating the learning process. Their model captures the interplay between images and videos, explaining video dynamics with a single image and residual information, and supports controlled video synthesis by incorporating additional factors such as motion direction. However, this kind of stochastic video generation can only handle short dynamic patterns in the distribution. Ni *et al.* [132] proposed a method for cI2V generation that synthesizes videos from a single image and a given condition, such as an action label. They introduced Latent flow diffusion models (LFDM), which generate an optical flow sequence in the latent space to warp the initial image, thereby improving the preservation of spatial details and motion continuity. The method involves a two-stage training process: an unsupervised Latent flow auto-encoder (LFAE) to estimate latent optical flow between video frames, and a conditional 3D U-Net-based Diffusion model (DM) to produce temporally-coherent latent flow sequences based on the image and condition. During inference, the image is encoded to a latent map, the condition to an embedding, and the trained DM generates latent flow and occlusion map sequences. During inference, the image is encoded to a latent map, the condition to an embedding, and the trained DM generates latent flow and occlusion map sequences. These sequences warp the latent map to create a new latent map sequence, which is then decoded into video frames. The proposed method, with its decoupled training strategy and efficient operation in a low-dimensional latent flow space, reduces computational cost and complexity while ensuring easy adaptation to new domains.

Wang *et al.* [133] proposed a high-fidelity image-to-video generation method, named DreamVideo, which addresses issues of low fidelity and flickering in existing methods by employing a frame retention branch in a pre-trained video diffusion model. The approach preserves image details by perceiving the reference image through convolution layers and integrating these features with noisy latents. The model incorporates double-condition classifier-free guidance, allowing a single image to generate videos of different actions through varying

prompts, enhancing controllable video generation. DreamVideo's architecture includes a primary T2V model and an Image Retention block that infuses image control signals into the U-Net structure. During inference, the model combines text and image inputs to generate contextually consistent videos using CLIP text embeddings and a U-Net-based generative process. Additionally, the Two-Stage Inference method extends video length and creates varied content by using the final frame of one video as the initial frame for the next, showcasing the model's strong image retention and video generation capabilities. Zhang *et al.* [134] proposed I2VGen-XL, a method utilizing two stages of cascaded diffusion models to achieve high semantic consistency and spatiotemporal continuity in video synthesis. The approach addresses challenges in semantic accuracy, clarity, and continuity by decoupling semantic and qualitative factors, using static images as guidance. The base stage ensures semantic coherence and preserves content at low resolution with two hierarchical encoders—a fixed CLIP encoder for high-level semantics and a learnable content encoder for low-level details. The refinement stage enhances video resolution and refines details using a brief text input and a separate video diffusion model. Training involves initializing the base model with pre-trained SD2.1 parameters and moderated updates, while the refinement model undergoes high-resolution training and fine-tuning on high-quality videos. Inference employs a noising-denoising process and DDIM/DPM-solver++ to generate high-resolution videos from low-resolution outputs.

To create more controllable videos, various motion cues like predefined directions and action labels are used. Blattmann *et al.* [135] proposed an approach for generating videos from static images by learning natural object dynamics through local pixel manipulations. Their generative model learns from videos of moving objects without needing explicit information about physical manipulations and infers object dynamics in response to user interactions, understanding the relationships between different object parts. The goal is to predict object deformation over time from a static image and a local pixel shift, using two encoding functions: an object encoder for the current object state and an interaction encoder for the pixel shift. They utilize a hierarchical recurrent model to understand complex object dynamics, predicting a sequence of object states in response to the pixel shift. Object dynamics are modeled using a flexible prediction function based on Recurrent Neural Networks (RNN), with higher-order dynamics captured by introducing a hierarchy of RNN predictors operating on different spatial scales. The decoder generates individual image frames from the predicted object states using a hierarchical image-to-sequence UNet structure. Instead of ground-truth interactions, dense optical flow displacement maps are used to simulate training pokes, minimizing the perceptual distance between predicted and actual video frames. Training involves pretraining the encoders and decoder to reconstruct image frames, then refining the model to predict object states and synthesize video sequences. Their interactive I2V synthesis model allows users to specify the desired motion through the manual poking of a pixel.

In addition, Menapace *et al.* [136] proposed a novel framework for the Playable video generation (PVG) task, which generates videos from the first frame and a sequence of discrete actions. While the PVG task reduces user burden by not requiring detailed motion information, it struggles with generating videos involving complex motions. An unsupervised learning approach is adopted that allows users to control video generation by selecting discrete actions at each time step, similar to video games. The framework, named Clustering for Action Decomposition and DiscoverY (CADDY), learns semantically consistent actions and generates realistic videos based on user input using a self-supervised encoder-decoder architecture driven by a reconstruction loss on the generated video. CADDY discovers distinct actions via clustering during the generation process, employing an encoder-decoder with a discrete bottleneck layer to capture frame transitions without needing action label supervision or a predefined number of actions. The action network estimates action label posterior distributions by decomposing actions into discrete labels and continuous components, ensuring meaningful action labels by preventing direct encoding of environment changes in the variability embeddings.

Within the generation of dynamic videos from static images presents a trend very similar to the previous section, Text-to-Video Synthesis, where we can see how attention mechanisms, autoencoders and diffusion models stand out. As we can see, GANs are not as frequent as in synthetic image generation. This approach to video generation can raise more ethical concerns than the previous one, as it can use images of real people and generate videos that can potentially harm them; whereas in the previous section, it involves content generated completely from scratch.

3. Video-to-Video Synthesis

Video-to-video (V2V) synthesis is an advanced field focused on generating realistic video sequences by transforming or translating visual information from one video domain to another. The main goal is to create high-quality, temporally consistent videos that adhere to specific input conditions, such as text, pose, style, or semantic maps. Recent advancements in this area have introduced several techniques to enhance the quality, efficiency, and consistency of video synthesis, thus pushing the boundaries of what is possible in video generation. Wang *et al.* [137] proposed a three-stage framework for human pose transfer in videos, focusing on transferring dance poses from a source person in one video to a target person in another. The process begins with the extraction of frames and pose masks from both source and target videos. Subsequently, a model synthesizes frames of the target person in the desired dance pose, followed by a refinement phase to enhance the quality of these frames. The model comprises several key components, including pose extraction and normalization, a GAN-based synthesis using Cross-domain correspondence network (CoCosNet), and a coarse-to-fine strategy with two GANs for detailed face reconstruction and smooth frame sequences. Their approach involves visualizing keypoints to create pose skeleton labels, adjusting for differences in body proportions, learning the translation from pose domain to image domain, and matching features for coherent synthesis. Although their method outperforms existing approaches, it still encounters challenges with large pose variations and domain generalization, which suggests potential areas for future improvement.

Furthermore, Zhuo *et al.* [138] introduced Fast-Vid2Vid, a spatial-temporal compression framework designed to reduce computational costs and accelerate inference in Video-to-Video synthesis (Vid2Vid). While traditional Vid2Vid generates photorealistic videos from semantic maps, it suffers from high computational costs due to the network architecture and sequential data streams. Zhuo *et al.* addressed this by introducing Motion-aware inference (MAI) to compress the input data stream without altering network parameters and developing Spatial-temporal knowledge distillation (STKD) to transfer knowledge from a high-resolution teacher model to a low-resolution student model. Their approach incorporates Spatial knowledge distillation (Spatial KD) for generating high-resolution frames from low-resolution inputs and Temporal knowledge distillation (Temporal KD) to maintain temporal coherence in sparse video sequences. Additionally, they utilize a part-time student generator for sparse frame synthesis and a fast motion compensation method for interpolating intermediate frames, thereby reducing computational load while maintaining visual quality. Further advancing the field, Yang *et al.* [139] introduced a zero-shot text-guided video-to-video translation framework that adapts image

models for video applications. This framework is composed of key frame translation and full video translation. Key frames are generated using an adapted diffusion model with hierarchical cross-frame constraints to ensure coherence in shapes, textures, and colors. These frames are then propagated to the rest of the video using temporal-aware patch matching and frame blending, achieving both global style and local texture temporal consistency without requiring re-training or optimization. A key innovation of this approach is the use of optical flow for dense cross-frame constraints, ensuring consistency across different stages of diffusion sampling. However, the method's reliance on accurate optical flow can lead to artifacts if the flow is incorrect, and significant appearance changes may disrupt temporal consistency, limiting the ability to create unseen content without user intervention.

Following the trend of previous researchers but focusing on zero-shot techniques, Wang *et al.* [140] presented vid2vid-zero, a zero-shot video editing method that leverages pre-trained image diffusion models without requiring video-specific training. Their method introduces a null-text inversion module for text-to-video alignment, a cross-frame modeling module for temporal consistency, and a spatial regularization module to preserve the fidelity of the original video. Vid2vid-zero addresses the issue of flickering in frame-wise image editing by ensuring temporal consistency through a Spatial-temporal attention (ST-Attn) mechanism, which balances bi-directional temporal information and spatial alignment using pre-trained diffusion models. While effective in video editing tasks, the method's reliance on pre-trained image models limits its capacity to edit actions in videos due to the absence of temporal and motion priors. Expanding on the idea of zero-shot video editing, Qi *et al.* [141] proposed FateZero, a zero-shot text-based editing method for real-world videos that does not require per-prompt training or user-specific masks. To achieve consistent video editing, FateZero utilizes techniques based on pre-trained models, capturing intermediate attention maps during DDIM inversion to retain structural and motion information and fusing these maps during editing. A blending mask, derived from cross-attention features, minimizes semantic leakage, while the reformed self-attention mechanism in the denoising UNet enhances frame consistency. Despite its impressive performance, FateZero faces challenges in generating entirely new motions or significantly altering shapes.

Other authors have opted for the use of diffusion models, due to their performance in similar tasks. Molad *et al.* [142] proposed Dreamix, a text-driven video editing method that uses a text-conditioned video diffusion model (VDM). Dreamix preserves the original video's fidelity by initializing with a degraded version of the input video and then fine-tuning the model. This mixed fine-tuning technique enhances motion editability by incorporating individual frames with masked temporal attention. Dreamix achieves text-guided video editing by inverting corruptions, downsampling the input video, corrupting it with noise, and then upscaling it using cascaded diffusion models aligned with the text prompt. This approach effectively preserves low-resolution details while synthesizing high-resolution outputs. Focusing on motion guidance, Hu *et al.* [143] introduced VideoControlNet, a motion-guided video-to-video translation framework using a diffusion model with ControlNet. Inspired by video codecs, VideoControlNet leverages motion information to maintain content consistency and prevent redundant regeneration. The first frame (I-frame) is generated using the diffusion model with ControlNet, mirroring the structure of the input frame. Key frames (P-frames) are then generated using the motion-guided P-frame generation (MgPG) module, which employs motion information for consistency and inpaints occluded areas using the diffusion model. The remaining frames (B-frames) are efficiently interpolated using the motion-guided B-frame interpolation (MgBI) module. This framework produces high-quality, consistent videos by utilizing advanced inpainting methods alongside motion information.

Adding to the discussion of temporal consistency, Liang *et al.* [144] introduced FlowVid, a V2V synthesis framework that ensures temporal consistency across frames by leveraging spatial conditions and temporal optical flow clues from the source video. Unlike previous methods, FlowVid uses optical flow as a supplementary reference to handle imperfections in flow estimation. The model warps optical flow from the first frame and uses it in a diffusion model, enabling the propagation of edits made to the first frame throughout subsequent frames. FlowVid extends the U-Net architecture to include a temporal dimension and is trained using joint spatial-temporal conditions, such as depth maps and flow-warped videos, to maintain frame consistency. During generation, the model edits the first frame with prevalent Image-to-image (I2I) models and propagates these edits using a trained model, incorporating global color calibration and self-attention feature integration to preserve structure and motion, thus achieving effective video synthesis with high temporal consistency. In a similar pursuit of enhancing temporal coherence, Wu *et al.* [145] proposed Fairy, a minimalist yet robust adaptation of image-editing diffusion models for video editing. Fairy improves temporal consistency and synthesis fidelity through anchor-based cross-frame attention, which propagates diffusion features across frames. To handle affine transformations, Fairy employs a unique data augmentation strategy, enhancing the model's equivariance and consistency. The anchor-based model samples K anchor frames to extract and propagate diffusion features, ensuring consistency by aligning similar semantic regions across frames. While Fairy excels in maintaining temporal consistency, its strong focus on this aspect reduces its accuracy in rendering dynamic visual effects, such as lightning or flames.

Lastly, several other methods offer significant contributions to the video-to-video synthesis domain. Ku *et al.* [146] proposed AnyV2V, see Fig. 9, a training-free video editing framework that simplifies video editing into two steps: editing the first frame with any image editing model and using an image-to-video generation model to create the edited video through temporal feature injection. AnyV2V is compatible with various image editing tools, allowing for diverse edits such as style transfer, subject-driven editing, and identity manipulation, without the need for fine-tuning. The framework uses DDIM inversion for structural guidance and feature injection to maintain consistency in appearance and motion, enabling accurate and flexible video editing. Additionally, it supports long video editing by handling videos beyond the training frame lengths of current I2V models, outperforming existing methods in user evaluations and standard metrics. Ouyang *et al.* [147] introduced I2VEdit, a video editing solution designed to extend the capabilities of image editing tools to videos. This approach achieves this by propagating single-frame edits throughout an entire video using a pre-trained Image-to-video model. Notably, I2VEdit adapts to the extent of edits, preserving visual and motion integrity while handling various types of edits, including global, local, and moderate shape changes. The method's core processes, coarse motion extraction and appearance refinement, play crucial roles in ensuring consistency. Coarse motion extraction captures basic motion patterns through a motion LoRA and employs skip-interval cross-attention to mitigate quality degradation in long videos.

Meanwhile, appearance refinement uses fine-grained attention matching for precise adjustments and incorporates Smooth area random perturbation (SARP) to enhance inversion sampling. To achieve its results, I2VEdit segments the source video into clips, processes each clip for motion and appearance consistency, and refines appearances using EDM [99] inversion and attention matching. Building on this, Ouyang *et al.* [148] further proposed Content deformation field (CoDeF), a novel video representation, emphasizing its application in Video-to-video translation. CoDeF introduces a canonical content field for static content aggregation and a temporal deformation field for recording

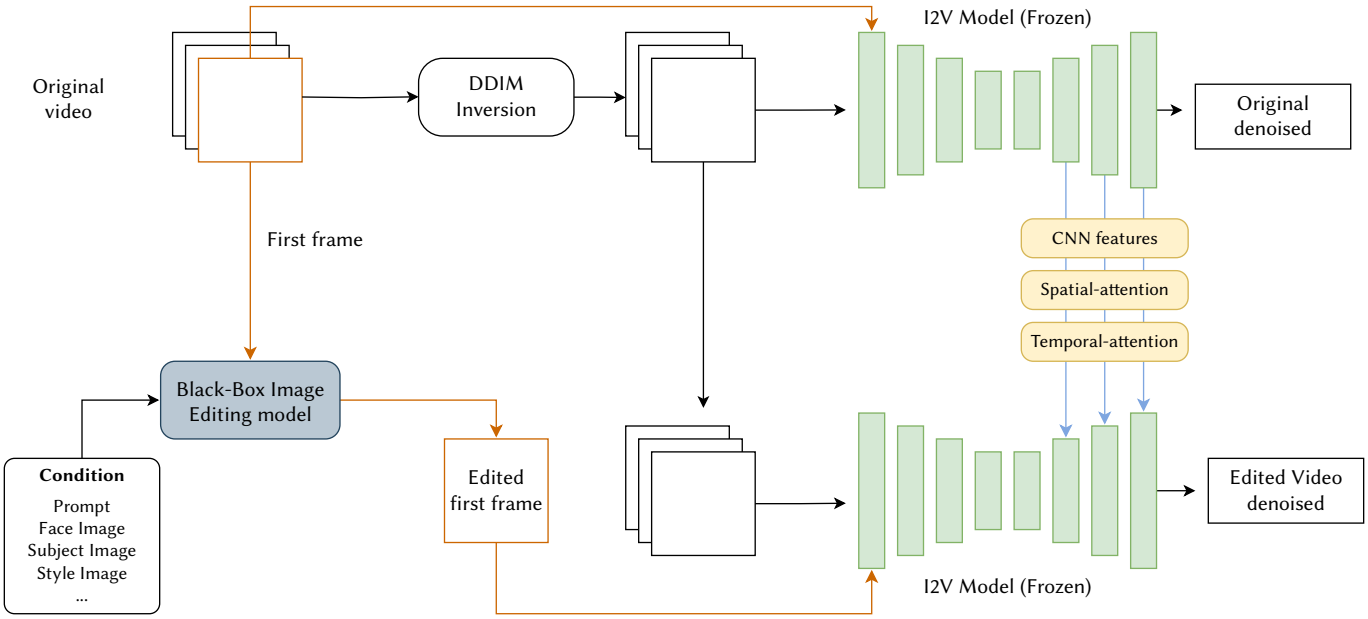


Fig. 9. AnyV2V [146] framework.

frame transformations. This approach optimizes the reconstruction of videos while preserving essential semantic details, such as object shapes. In the context of Video-to-video translation, CoDeF employs ControlNet on the canonical image, which significantly enhances temporal consistency and texture quality compared to state-of-the-art zero-shot video translations using generative models. By avoiding the need for time-intensive inference models, this process becomes more efficient. The canonical image, optimized through CoDeF, serves as a basis for applying image algorithms, ensuring consistent effect propagation across the entire video via the temporal deformation field.

A different approach to video editing with VideoSwap was presented by Gu *et al.* [149], focusing on customized video subject swapping. Unlike methods relying on dense correspondences, VideoSwap utilizes semantic point correspondences, allowing the replacement of the main subject in a video with a target subject of a different shape and identity, all while preserving the original background. The approach includes encoding the source video, applying DDIM inversion, and using semantic points to guide the subject's motion trajectory. The process also involves extracting and embedding semantic points, registering these points for motion guidance, and enabling user interactions to refine motion and shape alignment. Recently, Bai *et al.* [150] proposed UniEdit, a tuning-free framework for video motion and appearance editing. This framework leverages a pre-trained Text-to-video generator in an inversion-then-generation pipeline. UniEdit addresses content preservation by using temporal and spatial self-attention layers to encode inter-frame and intra-frame dependencies. Additionally, it introduces auxiliary reconstruction and motion-reference branches to inject the desired source and motion features into the main editing path. For content preservation, the auxiliary reconstruction branch injects attention features into the spatial self-attention layers. Motion injection, on the other hand, is achieved by guiding the main path with a motion-reference branch during denoising, utilizing temporal attention maps for alignment with the target prompt. In appearance editing, UniEdit maintains structural consistency by implementing spatial structure control while omitting the motion-reference branch. Despite its robust capabilities, UniEdit faces challenges, particularly when addressing motion and appearance editing simultaneously.

In this section we have analyzed the latest research related to video-to-video synthesis. Within this field we have seen how the most used technique is the diffusion model, as we have seen in other sections of this survey. This is in line with expectations due to all the attention they are receiving in recent years. However, we can also see that other methodologies such as GANs or attention mechanisms are also used. We have also noted that several papers use a zero-shot approach to address the problem.

4. Text-Image-to-Video Synthesis

Text-image-video synthesis (TI2V) is a growing field of research focused on generating dynamic video content from static images and text descriptions. Given a single image I and text prompt T, text-image-to-video generation aims to synthesize I new frames to yield a realistic video, $I^{=} \langle I^0, I^1, ..., I^M \rangle$ y starting from the given frame I^0 and satisfying the text description T . This field aims to bridge the gap between different modalities to create coherent and contextually accurate videos. Several approaches have been developed to address the challenges in this domain, ranging from aligning visual and textual information to ensuring temporal consistency and control over generated content. Hu *et al.* [151] proposed a novel video generation task called Text-Image-to-Video (TI2V) generation, which creates videos from a static image and a text description, focusing on controllable appearance and motion. They introduced the Motion Anchor-based video GEnerator (MAGE) to address key challenges such as aligning appearance and motion from different modalities and handling text description uncertainties. MAGE uses a Motion anchor (MA) structure to store aligned appearance-motion representations and incorporates explicit conditions and implicit randomness to enhance diversity and control. The framework employs a VQ-VAE encoder-decoder architecture for visual token representation and uses three-dimensional axial transformers to recursively generate frames. Training involves a supervised learning approach to approximate the conditional distribution of video frames based on the initial image and text. The motion anchor aligns text-described motion with visual features, ensuring consistent and diverse video output through auto-regressive frame generation.

Complementing this, Guo *et al.* [152] proposed AnimateDiff, a practical framework for animating personalized T2I models without requiring model-specific tuning. The core of the framework is a plug-and-play motion module, trained to learn transferable motion priors from real-world videos, which can be integrated into any

personalized T2I model. The training process involves three stages: fine-tuning a domain adapter to align with the target video dataset, introducing and optimizing a motion module for motion modeling, and using MotionLoRA, a lightweight fine-tuning technique, to adapt the pre-trained motion module to new motion patterns with minimal data and training cost. AnimateDiff effectively addresses the problem of animating personalized T2Is while preserving their visual quality and domain knowledge, demonstrating the adequacy of Transformer architecture for modeling motion priors and offering an efficient solution for users who desire specific motion effects without bearing the high costs of pre-training. In contrast, Yin *et al.* [153] proposed NUWA-XL, a novel "Diffusion over Diffusion" architecture for generating extremely long videos. Unlike traditional methods that generate videos sequentially, leading to inefficiencies and a training-inference gap, NUWA-XL uses a "coarse-to-fine" process where a global diffusion model generates keyframes and local models fill in between, allowing parallel generation. The architecture incorporates Temporal KLVAE to compress videos into low-dimensional latent representations and Mask temporal diffusion (MTD) to handle both global and local diffusion processes using masked frames. Although NUWA-XL is currently validated on cartoon data due to the lack of open-domain long video datasets, it shows promise in overcoming data challenges and improving efficiency, albeit requiring substantial GPU resources for parallel inference.

Esser *et al.* [154] proposed a structure and content-guided video diffusion model that edits videos based on user descriptions. They resolved conflicts between content and structure by training on monocular depth estimates with varying detail levels and introduced a novel guidance method for temporal consistency through joint video and image training. The approach extends latent diffusion models to video by incorporating temporal layers into a pre-trained image model, adding 1D convolutions and self-attentions to residual and transformer blocks. The encoder downsamples images to a latent code, improving efficiency, while depth maps and CLIP embeddings are used for structure and content conditioning, respectively. This approach allows full control over temporal, content, and structure consistency without requiring per-video training or pre-processing, showing improved temporal stability and user preference over related methods. Expanding on the concept of control, Yin *et al.* [155] proposed DragNUWA, an open-domain diffusion-based video generation model that integrates text, image, and trajectory inputs to provide fine-grained control over video content from semantic, spatial, and temporal perspectives. They address the limitations of current methods, which focus on only one type of control and struggle with complex trajectory handling, by introducing advanced trajectory modeling techniques: a Trajectory sampler (TS) for arbitrary trajectories, Multiscale fusion (MF) for controlling trajectories at different granularities, and an Adaptive training (AT) strategy for generating consistent videos. DragNUWA can generate realistic and contextually consistent videos by leveraging the combined inputs of text, images, and trajectories during both training and inference.

Further enhancing controllability, Wang *et al.* [156] proposed VideoComposer, a system for enhancing controllability in video synthesis through the use of temporal conditions like motion vectors. They introduced a Spatio-temporal condition encoder (STC-encoder) to integrate spatial and temporal dependencies, ensuring inter-frame consistency. The system decomposes videos into textual, spatial, and temporal conditions, and uses a latent diffusion model to recompose videos based on these inputs. Textual conditions provide coarse-grained visual content, while spatial conditions offer structural and stylistic guidance. Temporal conditions, including motion vectors and depth sequences, allow detailed control of temporal dynamics. Recently, Ni *et al.* [157] proposed TI2V-Zero, a zero-shot, tuning-free method for text-conditioned Image-to-video (TI2V) generation that leverages a pretrained T2V diffusion model. This approach avoids costly training, fine-tuning, or additional modules by using a "repeat-and-slide" strategy to condition video generation on a provided image, ensuring temporal continuity through a DDPM inversion strategy and resampling techniques. The method uses a 3D-UNet-based denoising network and modulates the reverse denoising process to generate videos frame-by-frame, preserving visual coherence and consistency, thus enabling the synthesis of long videos while maintaining high visual quality.

In this section where we have analyzed the techniques to generate videos from static images and textual descriptions, we have seen again a main focus, which are the diffusion models, i.e. a trend is observed, which seems to show that it will be the most used technique in the coming years. In addition, we also continue to observe other approaches such as attention mechanisms or autoencoders. The greatest danger of this set of techniques, like the previous one, is that they can use images of people to create complete videos, which can cause serious damage. However, not all applications of these techniques are negative.

5. Multi-Modal Video Generation

Multi-Modal Video Generation (MMVG) refers to a versatile field in which video content is synthesized based on different forms of input, such as text, images, or existing videos. Although models like Sora and Genie can accept various types of input, they typically process one modality at a time—either generating videos from text descriptions, animating static images, or transforming existing video footage. These approaches leverages the strengths of different data modalities to produce highly realistic and contextually coherent videos. The core objective of MMVG is to create coherent, high-fidelity, temporal consistent videos by leveraging the strengths of each input type. Recent advancements in this field have led to the development of sophisticated models capable of interpreting and synthesizing complex scenes by concurrently analyzing textual descriptions, visual cues, and pre-existing video footage. These models push the boundaries of video generation, offering versatile applications in content creation, entertainment, and beyond.

More recently, *OpenAI* [6] introduced Sora, a diffusion model that represents a significant advancement in T2V generation by training a model from scratch rather than fine-tuning pre-trained models. Drawing from transformer architecture scalability, Sora replaces the conventional U-Net with a transformer-based structure, effectively managing large-scale video data for complex generative tasks. Sora can generate high-fidelity videos up to a minute long, maintaining visual quality and narrative consistency across multiple shots. It leverages a patch-based approach, turning visual data into spacetime patches, which enhances its ability to handle videos and images of varying durations, resolutions, and aspect ratios. Sora excels in linguistic comprehension, accurately following detailed prompts to generate coherent video content. However, it faces challenges in rendering realistic interactions and comprehending complex scenes with multiple active elements. Despite these limitations, Sora's capabilities in video-to-video editing, image animation, and extending generated videos mark a significant step toward building general-purpose simulators of the physical world. Bruce *et al.* [7] introduced Genie, a generative interactive environment model trained unsupervised from unlabelled Internet videos. Genie uses spatiotemporal transformers, a novel video tokenizer, and a causal action model to create diverse, action-controllable virtual worlds from various inputs such as text, images, and sketches. It generates video frames autoregressively, enabling interaction on a frame-by-frame basis without ground-truth action labels.

TABLE IV. Comprehensive Overview of a Few Synthetic Video Generation Techniques

Models	Year	Technique	Target Outcome	Data Used	Open Source
Make-A-Video [91]	2023	Transformer-based	Text-to-video synthesis	Various	No
Video Diffusion [89]	2023	Diffusion-based	High-quality video synthesis	Video datasets	No
VideoPoet [5]	2023	Transformer-based	Generate poetic video narratives	Web-collected dataset	No
Godiva [104]	2023	GAN-based	Generate dynamic video content	High-resolution video datasets	No
CogVideo [106]	2023	Transformer-based	Extend CogView into video	Diverse text and video datasets	Yes
NUWA [107]	2023	Transformer-based	Synthesize coherent video clips	Diverse content from web datasets	No
NUWA-Infinity [108]	2023	Transformer-based	Generate endless video streams	Extended NUWA dataset	No
VideoGPT [109]	2023	GPT-based	Utilize GPT architecture	Various video datasets	Yes
Video LDMs [110]	2024	Latent Diffusion Models	Implement latent space techniques	Various	No
Text-to-Video (T2V) [158]	2023	Transformer-based	Synthesize video from static images	Diverse image and video datasets	No
ModelScope Text-to-Video [113]	2024	Transformer-based	Scalable text-to-video model	Large-scale web-collected video datasets	Yes
W.A.L.T [114]	2023	Diffusion Models	Enhance video synthesis	Various	No
C-ViViT [115]	2023	VAE-based	Create detailed videos from categories	Category-labeled video datasets	No
Text2Video-Zero [118]	2023	Zero-Shot Learning	Generate videos without explicit training	General video datasets	Yes
ART•V [119]	2024	AI Rendered Textures	Artistic video creation	Artistic style datasets	No
BIVDiff [120]	2023	Bi-directional Diffusion	Bidirectional control over video generation	Various	Yes
Simple Diffusion Adapter [116]	2024	Diffusion Models	Simplify diffusion processes	Various	Yes
HiGen [121]	2024	Hierarchical Generation	Layered approach to video scenes	Multi-layer video datasets	Yes

TABLE V. Overview of Techniques for Detecting AI-Generated Videos

Authors	Year	Technique	Target Outcome	Data Used	Open Source
Vahdati *et al.* [159]	2024	Synthetic video detection by forensic trace analysis	Detect AI-generated synthetic videos	Synth-vid-detect	No
He *et al.* [160]	2024	Temporal defects analysis	Identify temporal defects in AI-generated videos	ExposingAI-Video	No
Chen *et al.* [162]	2024	Detail Mamba for spatial-temporal artifacts detection	Enhance detection of AI-generated videos	GenVideo	Yes
Bai *et al.* [163]	2024	Spatio-temporal CNN analysis	Detect AI-generated videos using motion discrepancies	GVD	Yes
Ma *et al.* [164]	2024	Temporal artifact focus	Focus on temporal artifacts in video detection	GVF	Yes
Ji *et al.* [165]	2024	Dual-Branch 3D Transformer	Integrate motion and visual appearance for fake video detection	GenVidDet	No
Liu *et al.* [167]	2024	Diffusion-generated video detection	Capture spatial and temporal features in RGB frames and DIRE values	TOINR	No

As we can see, this section, multimodal video generation, is the least explored of all the approaches analyzed, see Table IV, and possibly the most complex, since we not only have to generate the visual part of the videos, but also the audio. In addition, we must ensure that both are matched and do not generate easily detectable artifacts. The techniques analyzed in this field are diffusion models and transformers. Possibly this area will be explored in more detail in the coming years.

B. Detection of AI-Generated Videos

In the rapidly evolving landscape of Generative AI (Gen AI), significant progress has been made in developing techniques to detect AI-generated synthetic images. Given that a video can be viewed as a sequence of images, one might reasonably expect that synthetic image detectors would also be effective at identifying AI-generated synthetic videos. Surprisingly, Vahdati *et al.* [159] reveal that current synthetic image detectors fail to reliably detect synthetic videos. Their study demonstrates that the forensic traces left by synthetic video generators are markedly different from those produced by image generators. This issue is not due to the degradation effects of H.264 compression but rather to the distinct characteristics of video generation. Therefore, their findings underscore the urgent need for detection methods tailored specifically to synthetic video content. Table V provides an overview of the techniques used for detecting AI-generated videos, highlighting key approaches and their application to various datasets. Despite the growing concerns, research into detecting synthetic videos has been relatively limited. Video generation technology is still in its early stages compared to image generation, and as a result, fewer detection methods are available. However, recent efforts have started to address this gap (see Fig. 10).

One early approach comes from, He *et al.* [160] who proposed a novel detection method for identifying AI-generated videos by analyzing temporal defects at both local and global levels. The method is based on the assumption that AI-generated videos exhibit different temporal dependencies compared to real videos due to their distinct capturing

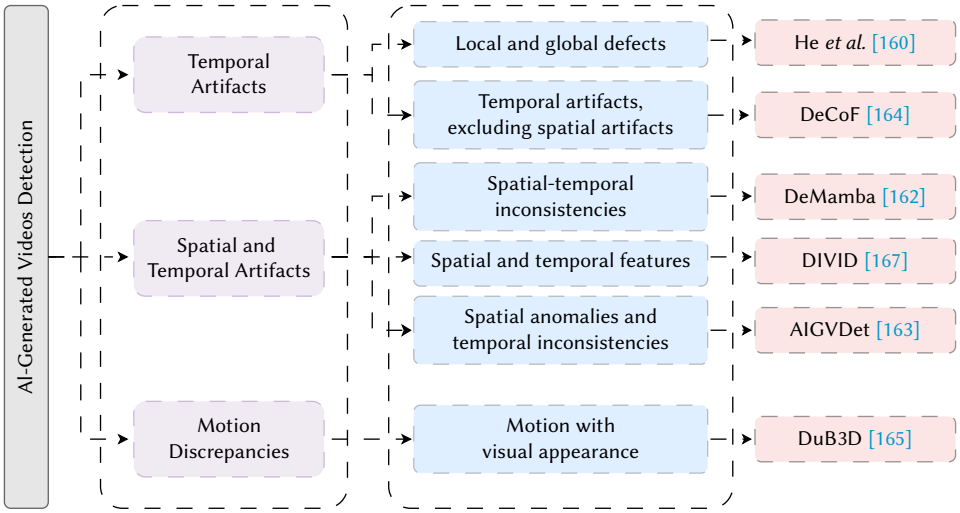


Fig. 10. AI-Generated videos detection methods overview.

and generation processes. Real videos, which are captured by cameras, have high temporal redundancy, whereas AI-generated videos control frame continuity in the latent space, leading to defects at different spatio-temporal scales. To address local motion information, the method uses a frame predictor trained on real videos to measure inter-frame motion predictability. Fake videos show larger prediction errors because they have less temporal redundancy. Temporal aggregation is employed to maintain long-range information and reduce the impact of diverse spatio-temporal details. The aggregated error map is then processed by a 2D encoder to obtain local motion features. For global appearance variation, the method extracts visual features using a pre-trained BEiT v2 [161] image encoder. These features are fed into a transformer to model temporal variations, identifying abnormal appearance changes across frames. Finally, a channel attention-based fusion module combines the local motion and global appearance features to enhance detection reliability. This module adjusts channel significance to extract more generalized forensic clues.

Furthermore, Chen *et al.* [162] proposed a plug-and-play module named Detail Mamba (DeMamba), designed to enhance the detection of AI-generated videos by identifying spatial and temporal artifacts. DeMamba builds upon the Mamba framework to explore both local and global spatial-temporal inconsistencies, addressing the limitation of models that consider only one aspect, either spatial or temporal. Using vision encoders like CLIP and XCLIP, it encodes video frames into a sequence of features, groups them spatially, and applies the DeMamba module to model intra-group consistency. Aggregated features from different groups help determine video authenticity. The DeMamba module introduces a novel approach to spatial consolidation by splitting features into zones along height and width, performing a 3D scan for spatial-temporal input. Unlike previous mechanisms, DeMamba's continuous scan aligns spatial tokens sequentially, enhancing the model's ability to capture complex relationships. For classification, DeMamba averages input features to obtain global features and pools processed features into local features, concatenating them with the global ones for classification via a simple MLP, ensuring robust video authenticity detection.

Based on the assumption that low-quality videos show abnormal textures and physical rule violations, while high-quality videos, indistinguishable to the naked eye, often manifest temporal discontinuities in optical flow maps, Bai *et al.* [163] proposed an effective AI-generated video detection (AIGVDet) scheme by capturing forensic traces with a two-branch spatio-temporal Convolutional Neural Network (CNN). This scheme employs two ResNet sub-detectors to identify anomalies in the spatial and optical flow domains. The spatial detector examines the abnormality of spatial pixel distributions within single RGB frames, while the optical flow detector captures temporal inconsistencies via optical flow. The model uses RGB frames and optical flow maps as inputs, with the two-branch ResNet50 encoder detecting abnormalities and a decision-level fusion binary classifier combining this information for the final prediction. AIGVDet effectively leverages motion discrepancies for comprehensive spatio-temporal analysis to detect AI-generated videos. Ma *et al.* [164] found that detectors based on spatial artifacts lack generalizability. Hence, they proposed DeCoF, a detection model that focuses on temporal artifacts and eliminates the impact of spatial artifacts during feature learning. DeCoF is the first method to use temporal artifacts by decoupling them from spatial artifacts, mapping video frames to a feature space where inter-feature distance is inversely correlated with image similarity, and detecting anomalies from inter-frame inconsistency. The method reduces computational complexity and memory requirements, needing only to learn anomalies between features. However, DeCoF may experience significant performance degradation or be inapplicable in the face of tampered video, such as Deepfake and malicious editing.

Traditional video detection models often overlook specific characteristics of downstream tasks, particularly in fake video detection where motion discrepancies between real and generated videos are significant, as generators tend to excel in appearance modeling but struggle with accurate motion representation. Ji *et al.* [165] proposed the Dual-Branch 3D Transformer (DuB3D) to address this issue by integrating motion information with visual appearance using a dual-branch architecture that fuses raw spatio-temporal data and optical flow. The spatial-temporal branch processes original frames to capture spatial-temporal information and identify anomalies, while the optical flow branch uses GMFlow [166] to estimate and capture motion information, and these features are combined using a Multi-layer perceptron (MLP) for classification. Built on the Video Swin Transformer backbone, DuB3D effectively enhances fake video detection by emphasizing motion modeling and demonstrating strong generalization across various video types. More recently, Liu *et al.* [167] proposed a novel approach for DIffusion-generated VIdeo Detection (DIVID). DIVID uses CNN+LSTM architectures to capture both spatial and temporal features in RGB frames and DIRE values. Initially, the CNN is fine-tuned on original RGB frames and DIRE values, followed by training the LSTM network based on the CNN's feature extraction. This two-phase training enhances detection accuracy for both in-

TABLE VI. AI-Generated Image Detection Datasets

Dataset	Year	Content	Real Source	Generator	#Real	#Generated	Available
LSUN Bed [168]	2022	Bedroom	LSUN	GAN/DM	420,000	510,000	✓
DFF [169]	2023	Face	IMDB-WIKI	DM	30,000	90,000	✓
RealFaces [170]	2023	Face	-	DM	-	25,800	✓
DiffusionForensics [79]	2023	General	LSUN ImageNet	DM	134,000	481,200	✓
Synthbuster [76]	2023	General	Raise-1k	DM	-	9,000	✓
DDDB [171]	2023	Art	LAION-5B	DM	64,479	73,411	✓
DE-FAKE [172]	2023	General	MSCOCO Flickr30k	DM	-	191 946	×
AI-Gen [173]	2023	General	ALASKA	DM	20,000	40,000	×
ArtiFact [174]	2023	General	Various sources including AFHQ, CelebAHQ, COCO, etc.	GAN/DM	964,989	1,531,749	✓
AutoSplice [175]	2023	General	Visual News	DM	2,273	3,621	✓
HiFi-IFDL [176]	2023	General	Various sources including AFHQ, CelebAHQ, LSUN, Youtube face etc.	GAN/DM	~ 600,000	1,300,000	✓
M3DSYNTH [177]	2023	CT	LIDC-IDRI	GAN/DM	1,018	8,577	✓
DIF [74]	2023	General	Laion-5B	GAN/DM	168,600	168,600	✓
DGM[4] [178]	2023	General	News: The Guardian, BBC, USA TODAY, Washington Post	GAN/DM	77,426	152,574	✓
COCOFake [179]	2023	General	COCO	DM	~ 1,200,000	~ 1,200,000	✓
DiFF [180]	2024	Face	VoxCeleb2 CelebA	DM	23,661	537,466	✓
CIFAKE [181]	2024	General	CIFAR-10	DM	60,000	60,000	✓
GenImage [182]	2024	General	ImageNet	GAN/DM	1,331,167	1,350,000	✓
Fake2M [183]	2024	General	CC3M	GAN/DM	-	2,300,000	✓
WildFake [184]	2024	General	Various sources including COCO, FFHQ Laion-5B, etc.	GAN/DM	1,013,446	2,680,867	×

domain and out-domain videos. Diffusion Reconstruction Error (DIRE) is calculated as the absolute difference between an original image and its reconstructed version from a pre-trained diffusion model, capturing signals of diffusion-generated images. By training the CNN+LSTM with DIRE and RGB frame features, DIVID improves detection accuracy for AI-generated videos.

Detecting AI-generated videos is an emerging challenge, distinct from synthetic image detection due to unique forensic traces in video content. While promising methods have begun to address this gap, leveraging spatio-temporal analysis and novel fusion techniques, the field is still evolving, see Table V. Continued innovation is essential to stay ahead of rapidly advancing video generation technologies.

V. Datasets

One of the most important aspects of DL model development is the availability of quality datasets. These datasets have to have some fundamental properties to be able to create robust models: to be representative, intra-class variability, balance between classes and a minimum quality. This will allow us to create suitable new generative and detection models. In this section we will focus on image and video datasets generated with AI.

The development of AI-generated images relies heavily on the availability of diverse and comprehensive datasets. These datasets provide the essential training material for models to learn from, enabling them to generate realistic and varied images. Ranging from large-scale collections of image-text pairs to datasets specifically designed for detecting synthetic content, these resources play a pivotal role in advancing the field. Regarding detection, we need representative and varied datasets that include different generation techniques and models. This will allow the development of robust models capable of being applied in real situations.

A. Image Datasets

In this section, we highlight some of the key image datasets that have significantly contributed to state-of-the-art AI-generated imagery. These datasets not only differ in size and content but also cater to various research needs, from general-purpose image generation to specialized tasks like AI-generated images detection and multimodal learning. For a detailed comparison, refer to Table VI, which summarizes the features and scope of these datasets.

Conceptual Captions 12M (CC12M) [185] is a large-scale dataset of 12.4 million image-text pairs derived from the Conceptual Captions 3M (CC3M) dataset [186]. CC12M was created by relaxing some of the filters used in CC3M to increase the recall of potentially useful image-alt-text pairs. The relaxed filters allow for more diverse and extensive data, though this results in a slight drop in precision. Unlike CC3M, CC12M does not perform hypernymization or digit substitution, except for substituting person names to protect privacy. This dataset's larger scale and diversity make it well-suited for vision-and-language pre-training tasks.

WIT [187] introduced to facilitate multimodal, multilingual learning, contains 37.5 million entity-rich image-text examples and 11.5 million unique images across 108 Wikipedia languages. It serves as a pre-training dataset for multimodal models, particularly useful for tasks like image-text retrieval. WIT stands out due to its large size, multilingual nature with over 100 languages, diverse concepts, and a challenging real-world test set. It combines high-quality image-text pairs from curated datasets like Flickr30K and MS-COCO with the scalability of extractive datasets. WIT's creation involved filtering

low-information associations and ensuring image quality. The dataset provides multiple text types per image (reference, attribution, and alt-text), offers extensive cross-lingual text pairs, and supports contextual understanding with 120 million contextual texts.

RedCaps [188] is a large-scale dataset introduced in 2021, consisting of 12 million image-text pairs collected from Reddit. This dataset includes images and captions depicting a variety of objects and scenes, sourced from a manually curated set of subreddits to ensure diverse yet focused content. The data collection process involves three steps: subreddit selection, image post filtering, and caption cleaning. Images are primarily photographs from 350 selected subreddits, excluding any NSFW, banned, or quarantined content. Filtering techniques are used to maintain high-quality captions and mitigate privacy and harmful stereotypes, resulting in a robust and extensive dataset.

Laion-5b [189] is a large-scale vision-language dataset derived from Common Crawl, containing nearly 6 billion image-text pairs. Images with alt-text were extracted and processed to remove low-quality and malicious content. Filtering based on cosine similarity with OpenAI's ViT-B/32 CLIP model reduced the dataset size significantly. The dataset is divided into three subsets: 2.32 billion English pairs, 2.26 billion multilingual pairs, and 1.27 billion pairs with undetected languages. Metadata includes image URLs, text, dimensions, similarity scores, and NSFW tags.

DiffusionDB [190] is the first large-scale prompt dataset totaling 6.5TB, containing 14 million images generated by Stable Diffusion using 1.8 million unique prompts. Constructed by collecting images shared on the Stable Diffusion public Discord server. Most prompts are between 6 to 12 tokens long, with a significant spike at 75 tokens, indicating many users exceed the model's limit. 98.3% of the prompts are in English, with the rest covering 34 other languages. DiffusionDB provides unique research opportunities in prompt engineering, explaining large generative models, and detecting deepfakes, serving as an important resource for studying prompts in text-to-image generation and designing next-generation human-AI interaction tools.

DiffusionForensics [79] is a dataset designed for evaluating diffusion-generated image detectors. It includes 42,000 real images from LSUN-Bedroom, 50,000 from ImageNet, and 42,000 from CelebA-HQ. Generated images are produced by various models, with unconditional models like ADM, DDPM, iDDPM, and PNDM generating 42,000 images each from LSUN-Bedroom. Text-to-image models LDM, SD-v1, SD-v2, and VQ-Diffusion also generate 42,000 images each, while IF, DALLE-2, and Midjourney produce fewer images. For ImageNet, 50,000 images each are generated by a conditional model ADM and a text-to-image model SD-v1. CelebA-HQ includes 42,000 images generated by SD-v2 and smaller sets by IF, DALLE-2, and Midjourney.

LSUN Bedroom [168] dataset contains images center-cropped to 256×256 pixels. Samples are either downloaded or generated using code and pre-trained models from original publications. The dataset includes samples from ten models (e.g. ProGAN, Diff-StyleGAN2, Diff-ProjectedGAN, DDPM, IDDPM,LDM). For each model, 51,000 images were sampled, and the real part is sourced from lsun bedroom dataset [191].

DeepFakeFace (DFF) [169] is a dataset designed to evaluate deepfake detectors, featuring 120,000 images, with 30,000 real images sourced from the IMDB-WIKI dataset and 90,000 fake images. To generate these fake images, three models were used: Stable Diffusion v1.5, Stable Diffusion Inpainting, and InsightFace, each producing 30,000 images. The dataset includes high-resolution images of 512 × 512 pixels. Real images were matched by gender and age, using prompts like "name, celebrity, age" for generation. Discrepancies in facial bounding boxes were corrected using the RetinaFace detector to ensure accuracy before generating deepfakes.

RealFaces [170] consists of 25,800 images generated using Stable Diffusion, incorporating prompts for photorealistic human faces. It includes 431 images filtered by an NSFW filter, mainly depicting women and young people.

Deepart Detection Database (DDDB) [171] is designed for detecting deepfake art. It includes high-quality conventional art from LAION-5B and deepfake art from models like Stable Diffusion, DALL-E 2, Imagen, Midjourney, and Parti. Conart images are sourced from LAION-5B, while deeparts are generated using state-of-the-art models or collected from social media. DDDB consists of 64,479 conventional art images (conart) and 73,411 deepfake art images (deepart). It supports research in deepart detection, continuously updating to incorporate new deeparts and addressing privacy and storage constraints.

SynthBuster [76]. Due to the scarcity of diffusion model-generated images, SynthBuster addresses this by providing a new dataset with images from models like Stable Diffusion 1.3, 1.4, 2, and XL, Midjourney, Adobe Firefly, and DALL·E 2 and 3. While synthetic images are generated from text, SynthBuster uses the existing Raise-1k database of real images, which is a varied subset of the Raise [192] dataset, as a guideline for the generated image. Original images are not used as prompts to try to recreate or modify a similar image. They are only used as a guideline to create the new prompt for the presentation, to ensure that the resulting image is broadly in the same category as the original image. For each of the 1000 images, descriptions are generated using the Midjourney descriptor [3] and CLIP Interrogator [193]. Then, these descriptions were used as the basis for manually writing a text prompt to generate a photo-realistic image loosely based on the original image.

DE-FAKE [172] is designed for detecting AI-generated images. Real images are sourced from the MSCOCO and Flickr30k datasets. To create a corresponding set of fake images, prompts from these real images were used to generate 191,946 synthetic images through four different image generation models: Stable Diffusion, Latent Diffusion, GLIDE, and DALLE-2.

AI-Gen [173] dataset consists of 20,000 uncompressed 256 × 256 PG images from the ALASKA [194] database, which are used to construct the T2I dataset. Specific spots and objects are extracted from these Photographs (PG) images, and 5,000 prompts are generated with ChatGPT. Two AI systems, DALL·E2 [195] and DreamStudio, are used to generate four images per prompt, creating two databases: DALL·E2 [195] and DreamStudio [196]. Each database contains 20,000 Photographs (PG) images and corresponding T2I images. The images are resized to 256 × 256, 128 × 128, and 64 × 64, and JPEG compression is applied with a quality factor between 75 and 95. The datasets are divided into training (12,000 pairs), validation (3,000 pairs), and testing (5,000 pairs).

AutoSplice [175] is a image dataset containing 5,894 manipulated and authentic images, designed to aid in developing generalized detection methods. The dataset consists of 3,621 images generated by locally or globally manipulating real-world image-caption pairs from the Visual News dataset. The DALL-E2 generative model was used to create synthetic images based on text inputs. AutoSplice construction involved pre-processing with object detection and text parsing, human annotations to select and modify object descriptions, and post-processing to filter out images with visual artifacts. The final dataset includes 3,621 high-quality manipulated images and 2,273 authentic images, with versions in both lossless and gently lossy JPEG compression formats.

ArtiFact [174] is a large-scale dataset designed to evaluate the generalizability and robustness of synthetic image detectors by incorporating diverse generators, object categories, and real-world

impairments. It includes 2,496,738 images, with 964,989 real and 1,531,749 fake images. The dataset covers multiple categories such as Human/Human Faces, Animal/Animal Faces, Places, Vehicles, and Art, sourced from 8 source datasets (e.g., COCO, ImageNet, AFHQ, Landscape) . It features images synthesized by 25 distinct methods, including 13 GANs (e.g., StyleGAN3, StyleGAN2, ProGAN), 7 Diffusion models (e.g., DDPM, Latent Diffusion, LaMA), and 5 other generators (e.g., CIPS, Palette). To ensure real-world applicability, images undergo impairments like random cropping, resizing, and JPEG compression according to IEEE VIP Cup 2022 standards.

CIFAKE [181] consists of 120,000 images, split evenly between real and synthetic images. The real images are taken from the CIFAR-10 [197] dataset, comprising 60,000 32x32 RGB images across ten classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, with 50,000 images used for training and 10,000 for testing. The synthetic images are generated using the CompVis Stable Diffusion model (version 1.4), which is trained on subsets of the LAION-5B [189] dataset. The generation process involves reverse diffusion from noise to create 6,000 images per class, mimicking the CIFAR-10 [197] dataset. Similar to the real images, 50,000 synthetic images are used for training and 10,000 for testing, with labels indicating their synthetic nature.

GenImage [182] is designed to evaluate detectors' ability to distinguish between AI-generated and real images. It includes 2,681,167 images, with 1,331,167 real images from ImageNet and 1,350,000 fake images generated using eight models: BigGAN, GLIDE, VQDM, Stable Diffusion V1.4, Stable Diffusion V1.5, ADM, Midjourney, and Wukong. The images are balanced across ImageNet's 1000 classes, with specific allocations for training and testing. Each model generates a nearly equal number of images per class, ensuring no overlap in real images. The dataset features high variability and realism, particularly in animals and plants, providing a robust basis for developing detection models.

Fake2M [183] dataset is a large-scale collection of over 2 million AI-generated images. These images are created using three different models: Stable Diffusion v1.5, IF, and StyleGAN3. The dataset aims to investigate whether models can distinguish AI-generated images from real ones.

DiFF [180] comprises over 500,000 images synthesized using thirteen distinct generation methods under four conditions, leveraging 30,000 textual and visual prompts to ensure high fidelity and semantic consistency. The dataset includes pristine images from 1,070 celebrities, curated from sources like VoxCeleb2 and CelebA, totaling 23,661 images. Prompts, derived from these pristine images, include original and modified textual prompts as well as visual prompts. The dataset covers four categories of diffusion models: Text-to-Image (T2I), Image-to-Image (I2I), Face Swapping (FS), and Face Editing (FE), employing methods like Midjourney, Stable Diffusion XL, DreamBooth, DiffFace, and others to generate the forged images.

WildFake [184] is designed to assess the generalizability and robustness of fake image detectors. Developed with diverse content from open-source websites and generative models, it provides a comprehensive set of high-quality fake images. It includes images from DMs, GANs, and other generators, with categories such as "Early" and "Latest" models. The dataset also features nine kinds of DMs generators and various fine-tuning strategies for SD-based generators. Images were collected using a generation pipeline from platforms like Civitai and Midjourney, ensuring a representative sample of real-world quality. Real images were sourced from datasets like COCO, FFHQ, and Laion-5B. WildFake contains 3,694,313 images, with 1,013,446 real and 2,680,867 fake images, split into training and testing sets in a 4:1 ratio.

B. Video Datasets

In this section, we review key video datasets that have been pivotal in advancing state-of-the-art AI models. These resources Offer diverse video-text pairs, high-resolution clips, and specialized content, each contributing uniquely to the progress of AI-driven video technology. For a detailed comparison, refer to Table VII, which summarizes the characteristics and scope of these datasets.

YT-Tem-180M [198] was collected from 6 million public YouTube videos, totaling 180 million clips, and annotated by ASR. It includes diverse content such as instructional lifestyle vlogs, and auto-suggested videos on topics. Videos were filtered to exclude those an English ASR track, over 20 minutes long, in 'ungrounded" categories, or with thumbnails to contain objects. Each video was split into segments of an image frame and corresponding spoken words, resulting in 180 million segments.

WebVid-2M [199] is a large-scale video-text pretraining dataset consisting of 2.5 million video-text pairs. The average length of each video is 18.0 seconds, and the average caption length is 12.0 words. The raw descriptions for each video are collected from the Alt-text HTML attribute associated with web images. This dataset was scraped from the web using a method similar to Google Conceptual Captions (CC3M), which includes over 10% of images that are video thumbnails. WebVid-2M captions are manually generated, well-formed sentences aligned with the video content, contrasting with the HowTo100M [105] dataset, which contains incomplete sentences from continuous narration that may not be temporally aligned with the video.

CATER-GEN-v1 [151] is a synthetic dataset set in a 3D environment, derived from CATER [210], featuring two objects (cone and snitch) and a large table plane. It includes four atomic actions: "rotate", "contain", "pick-place", and "slide", with each video containing one or two actions. Descriptions are generated using predefined templates, with a resolution of 256x256 pixels. The dataset includes 3,500 training pairs and 1,500 testing pairs.

CATER-GEN-v2 [151] is a more complex version of CATER-GEN-v1, containing 3 to 8 objects per video, each with randomly chosen attributes from five shapes, three sizes, nine colors, and two materials. The actions are the same as in CATER-GEN-v1, but descriptions are designed to create ambiguity by omitting certain attributes. The video resolution is 256x256 pixels, and the dataset includes 24,000 training pairs and 6,000 testing pairs.

Internvid [202] is a video-centric multimodal dataset created for large-scale video-language learning, featuring high temporal dynamics, diverse semantics, and strong video-text correlations. It includes 7 million YouTube video-text correlations. It includes 7 million YouTube videos with an average duration of 6.4 minutes, covering 16 topics. Videos were collected based on popularity and action-related queries, ensuring diversity by including various countries and languages. Each video is segmented into clips, resulting in 234 million clips from 2s to more than 30s duration, which were captioned using a multiscale method focusing on common objects and actions. InternVid emphasizes high resolution, with 85% of videos at 720P, and provides comprehensive multimodal data including audio, metadata, and subtitles. The dataset is notable for its action-oriented content, containing significantly more verbs compared to other datasets, and includes 7.1 million interleaved video-text data pairs for in-context learning.

FlintstonesHD [153] is a densely annotated long video dataset created to promote the development of long video generation. The dataset is built from the original Flintstones cartoon, containing 166 episodes with an average of 38,000 frames per episode, each at a resolution of 1440 × 1080 pixels. Unlike existing video datasets, FlintstonesHD addresses issues such as short video lengths, low

TABLE VII. Video Datasets. Datasets With Grey Background Are Used in A AI-generated Videos Detection

Dataset	Year	Source	Size	Domain	Resolution	Text	Avg len (sec)	Duration (hrs)	Unique Features
YT-Tem-180M [198]	2021	YouTube, HowTo100M	180M Videos 180M Text	Open	-	ASR	-	-	Filters to exclude non-English ASR and visuall «ungrounded» categories
WebVid-2M [199]	2021	Web	2.5M Videos 2.5M Text	Open	360p	Manual	18.0	13K	Manually generated captions, aligned with video content
WebVid-10M [199]	2021	Web	10M Videos 10M Text	Open	360p	Alt-Text	18.0	52K	Manually generated captions, aligned with video content
CATER-GEN-v1 [151]	2022	Synthetic 3D objects	5K Video 5K Text	Geometric	256p	Predefined template	-	-	Synthetic, simple scenes with atomic actions
CATER-GEN-v2 [151]	2022	Synthetic 3D objects	30K Video 30K Text	Geometric	256p	Predefined Template	-	-	Increased complexity with more objects and attributes
CelebV-HQ [200]	2022	Web	35,666 Videos	Face	512p	Manual	3 to 20	65	High-quality, detailed text descriptions
HD-VILA-100M [201]	2022	YouTube	103M Videos 103M Text	Open	720p	ASR	13.4	371.5K	High-quality alignment of videos and transcriptions
Internvid [202]	2023	YouTube	7.1M Videos 234M clips	Open	360p 512p 720p	Generated	11.7	760.3K	Action-oriented, diverse languages, and high video-text correlation
FlintstonesHD [153]	2023	Flintstones cartoon	166 episodes	Cartoon	1440x1080	Generated	-	-	Densely annotated for long video generation
Celebv-text [203]	2023	Web	70K Videos 1.4M Text	Face	512p+	Semi-Auto Generated	<5s	279	High-quality, detailed text descriptions
HD-VG-130M [204]	2023	YouTube	130M Videos 130M Text	Open	720p	Generated	~ 5.1	184K	High-definition, single-scene clips
Youku-mPLUG [205]	2023	Youku platform	10M Videos 10M Text	Open	-	-	54.2	150K	Focused on advancing Chinese multimodal LLMs
VidProM [206]	2024	Pika Discord	1.67M prompts 6.69M Videos	Open	-	Manual		-	Extensive prompts with semantic uniqueness
MiraData [207]	2024	YouTube, Videvo, Pixabay, Pexels HD-VILA-100M		Open	720p	Generated	72.1	16K	High visual quality, detailed captions
GenVideo [162]	2024	Kinetics-400 Youku-mPLUG MSR-VTT Video Gen Methods	~ 2.31M Videos	Open	-	Automatic	2 to 6		Balance of real and fake videos across diverse scenes
ExposingAI-Video [160]	2024	MSVD, Potat1 Ali-vilab,ZScope T2V-zero	2K Videos	Open	-	Automatic	-	-	H. 265 compression and quality degradation simulation
Synth-vid-detect [159]	2024	MIT, Video-ACID Gen Video Methods	18.75K Videos	Open	-	Automatic	-	-	H. 265 compression Out-of-distribution test set
GVD [163]	2024	GOT, Youtube_vos2 Gen Video Methods	-	Open	-	Automatic	-	-	Collection from various SOTA models
GVF [164]	2024	MSVD, MSR-VTT Gen Video Methods	964 Videos 964 Text	Open	-	Automatic	-	-	Diversity in forgery targets, scenes, and behaviors
GenVidDet [165]	2024	InternVid, HD-VG-130M Gen Video Methods	~2.66M Videos	Open	256p 512p 720p	Automatic	-	4442	Large-scale dataset cover diverse content
TOINR [167]	2024	VidVRD, SVD-XT YouTube SORA, Pika, GEN-2	~2.826K Videos	Open	-	Automatic	-	-	Out-domain testing with various generation tools
Panda-70m [208]	2024	HD-VILA-100M	70.8M Videos 70.8M Text	Open	720p	Automatic	8.5s	166.8K	High-quality captions with significant improvements in downstream tasks
VAST-27M [209]	2024	HD-VILA-100M	27M Videos 297M Text	Open	-	Generated	5 to 30 sec	-	Comprehensive with vision, audio, and omni-modality captions

resolution, and coarse annotations. The image captioning model GIT2 [211] was used to generate dense captions for each frame, with manual filtering to correct errors, thus providing detailed annotations that capture movement and story nuances. This dataset serves as a benchmark for improving long video generation.

Celebv-text [203] is a large-scale facial text-video dataset aimed at providing high-quality video samples with relevant, diverse text descriptions. Constructed through data collection and processing, data annotation, and semi-auto text generation, it features 70, 000 video clips totaling around 279 hours. Videos were sourced from the internet, using queries like human names and movie titles, excluding low-resolution and short clips, and processed to maintain high quality without upsampling or downsampling. Annotations include static attributes like general appearance and light conditions, and dynamic attributes like actions and emotions, with both automatic and manual methods used for accuracy. Texts were generated using a combination of manual descriptions and auto-generated templates based on common grammar structures, resulting in longer and more detailed text descriptions compared to other datasets. CelebV-Text surpasses existing datasets like MM-Vox [212] and CelebV-HQ [200] in scale, resolution, and text-video relevance, offering a comprehensive resource for facial video analysis.

VidProM [206] is a large-scale dataset for text-to-video diffusion models, collected from Pika Discord channels between July 2023 and February 2024.It includes 1,672,243 unique text-to-video prompts, embedded with 3072-dimensional embeddings using OpenAI's text-embedding-3-large API. The dataset includes NSFW probabilities assigned using the Detoxify model, with less than 0.5% of prompts flagged as potentially unsafe. It features 6.69 million videos generated by Pika, VideoCraft2, Text2Video-Zero, and ModelScope, involving significant computational resources. After filtering for semantic uniqueness, VidProM retains 1,038,805 unique prompts. Compared to DiffusionDB, VidProM has 40.6% more semantically unique prompts and supports longer, more complex prompts due to its advanced embedding model. VidProM includes videos generated by four state-of-the-art models, resulting in over 14 million seconds of video content. VidProM's extensive video content and complex prompts, requiring dynamic and temporal descriptions, make it a valuable resource for developing text-to-video generative models.

MiraData [207] is a large-scale text-video dataset with long durations and detailed structured captions. The dataset, finalized through a five-step process, sources videos from YouTube, Videvo, Pixabay, and Pexels to ensure diverse content and high visual quality. From YouTube, 156 high-quality channels were selected, resulting in 68K videos and 173K clips post-processing. Additional videos were sourced from HD-VILA-100M, Videvo (63K), Pixabay (43K), and Pexels (318K). Video clips were split and stitched using models like Qwen-VL-Chat and DINOv2, ensuring semantic coherence and content continuity. MiraData provides five versions of filtered data based on video color, aesthetic quality, motion strength, and NSFW content, with 788K to 9K clips. Captions were generated using GPT-4V, resulting in dense and structured descriptions with average lengths of 90 and 214 words respectively. MiraData surpasses previous datasets in visual quality and motion strength, making it ideal for text-to-video generation tasks.

GenVideo [162] is a large-scale dataset developed to evaluate the generalizability and robustness of AI-generated video detection models. The training set contains 2, 294, 594 video clips, including 1, 213, 511 real and 1, 081, 083 fake videos, while the testing set includes 19, 588 video clips, with 10, 000 real and 8, 588 fake videos. The dataset features high-quality fake videos sourced from open-source websites and various pre-trained models, covering a wide range of scenes such as landscapes, people, buildings, and objects. Video duration's range from 2 to 6 seconds, with diverse aspect ratios. Real videos are sourced from datasets like Kinetics-400, Youku-mPLUG, and MSR-VTT [213]. Fake videos are generated using diffusion-based models, auto-regressive models, and other methods such as VideoPoet, Emu, Sora, VideoCrafter, latent flow diffusion models, masked generative video transformer, and autoregressive models. Additionally, sources include external web scraping and service-based methods like the Pika website. This diverse and comprehensive collection aims to enhance the understanding and detection of AI-generated videos across numerous real-world contexts.

ExposingAI-Video [160] is composed of 1,000 natural videos sourced from the MSVD [214] dataset, paired with 1,000 fake videos generated using four advanced diffusion-based video generators, resulting in 96,000 fake frames. The dataset offers diverse content driven by text prompts, featuring rich motion information distinct from static images. It includes videos generated by models such as ali-vilab, zeroscope, potat1, and a zero-shot text-to-video model, each providing unique configurations. Additionally, the dataset incorporates three video post-processing operations—H.265 ABR compression, H.265 CRF compression, and Bit Error—to simulate quality degradation for robustness evaluation.

Synth-vid-detect [159] consists of both real and synthetic videos for training and evaluation. It includes 7,654 real videos for training, 784 for validation, and 1,661 for testing, sourced from the Moments in Time (MIT) [215] and Video-ACID [216] datasets. The synthetic videos, totaling 6,197 for training, 624 for validation, and 1,429 for testing, were generated using Luma, VideoCrafter-v1, CogVideo, and Stable Video Diffusion, with diverse scenes and activities represented. All videos were compressed using H.264 at a constant rate factor of 23. For testing, an exclusive set of prompts and videos was used to avoid overlap with the training data. Additionally, the dataset includes an out-of-distribution, test-only set of 401 synthetic videos generated by Sora, Pika, and VideoCrafter-v2.

Generated Video Dataset (GVD) [163] includes 11,618 video samples produced by 11 different state-of-the-art generator models. These models generate videos using either T2V or I2V techniques. The dataset was primarily collected from the Discord platform, where users share videos generated by various models. For training and validation, 550 T2V-generated videos from Moonvalley [217] and 550 real videos from the YouTube_vos2 [218] dataset were used. All generated videos not used in training and validation are designated for testing, with real test videos sourced from the GOT [219] dataset.

GeneratedVideoForensics (GVF) [164] dataset consists of 964 triples, each containing a real video, a corresponding text prompt, and a video generated by one of four different open-source text-to-video generation models: Text2Video-zero, ModelScopeT2V, ZeroScope, and Show-1. These models cover various forgery targets, scenes, behaviors, and actions, ensuring the dataset's diversity. The real videos and prompts were collected from MSVD [214] and MSR-VTT [213] datasets, with a focus on simulating realistic video distributions across spatial and temporal dimensions. It also includes vidoes from most popular commercial models like OpenAI's Sora, Pika, Gen-2 and Google's Veo.

GenVidDet [165] is a large-scale video dataset created for AI-generated video detection, comprising over 2.66 million clips with more than 4442 hours of content. It includes real videos sourced from the InternVid [202] and HD-VG-130M [204] datasets, totaling over 1.46 million clips, and AI-generated videos from the VidProM dataset using four different models, adding approximately 1.12 million clips. Additionally, new AI-generated videos were created using the latest models like Open-Sora, StreamingT2V and DynamiCrafter to enhance the dataset's diversity.

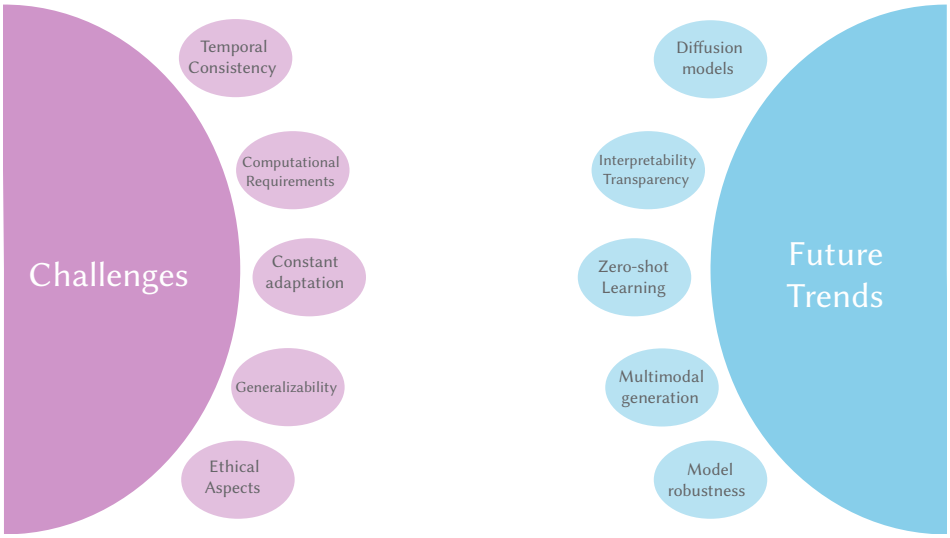


Fig. 11. Overview of trends and challenges in the generation and detection of AI-generated image and video samples.

Turns Out I'm Not Real (TOINR) [167] dataset was constructed to evaluate a method using public video generation tools, including Stable Video Diffusion (SVD), Pika, Gen-2, and SORA. The dataset includes 1,000 real video clips from the ImageNet Video Visual Relation Detection (VidVRD) [220] dataset and 1,000 fake video clips generated with SVD-XT [89]. It also comprises an additional real and fake clips for out-domain testing: 107 real (VidVRD) and 107 fake clips generated with Pika, 107 (VidVRD) real and 107 fake clips generated with Gen-2, and 207 real and 191 fake clips sourced from YouTube and SORA website.

HD-VILA-100M [201] is a high-resolution and diversified video-language dataset designed to overcome limitations in existing datasets. Introduced to aid tasks such as Text-to-video retrieval and video QA, it comprises 103 million video clip and sentence pairs from 3.3 million videos, totaling 371.5K hours. Sourced from diverse YouTube content, including professional channels like BBC Earth and National Geography, HD-VILA-100M emphasizes quality and alignment of videos and transcriptions. Only videos with subtitles and 720p resolution were included, resulting in a final set of 3.3 million videos, balanced across 15 categories. For video-text pairing, the dataset utilizes video transcriptions instead of manual annotations, offering richer information. Subtitles, often generated by ASR, were split into complete sentences using an off-the-shelf tool. Sentences were aligned with video clips using Dynamic Time Warping, producing pairs averaging 13.4 seconds in length and 32.5 words per sentence.

HD-VG-130M [204] is a large-scale dataset for Text-to-video generation, comprising 130 million text-video pairs from the open domain. Created to address limitations in existing datasets, it features high-definition (720p), widescreen, and watermark-free videos. Collected from YouTube, the videos were processed using PySceneDetect for scene detection, resulting in single-scene clips of less than 20 seconds each. Captions were generated using BLIP-2, ensuring that descriptions, typically around 10 words, are representative of the visual content. Covering 15 categories, HD-VG-130M provides diverse and high-quality data for training video generation models.

Youku-mPLUG [205] is the first Chinese video-language pretraining dataset, released in 2023 and collected from the Youku video-sharing platform. It comprises 10 million high-quality Chinese video-text pairs filtered from 400 million raw videos, covering 45 diverse categories with an average video length of 54.2 seconds. This dataset was created to advance Vision-language pre-training (VLP) and multimodal Large language models (LLMs) within the Chinese community. Strict criteria for safety, diversity, and quality were applied, involving multi-level risk detection to eliminate high-risk content and video fingerprinting to ensure a balanced distribution. Additionally, the dataset includes 0.3 million videos for downstream benchmarks, designed to assess video-text retrieval, video captioning, and video category classification tasks.

Panda-70m [208] is a large-scale video dataset created for video captioning, video and text retrieval, and text-driven video generation. It consists of 70 million high-resolution, semantically coherent video clips with captions. The dataset was developed from 3.8 million long videos collected from HD-VILA-100M [201]. To generate accurate captions, a two-stage semantics-aware splitting algorithm was used, followed by multiple cross-modality teacher models to predict candidate captions. A subset of 100,000 videos was manually annotated to fine-tune a retrieval model, which then selected the best captions for the entire dataset. Panda-70M addresses the challenge of collecting high-quality video-text data and shows significant improvements in downstream tasks. The dataset primarily contains vocal-intensive videos such as news, TV shows, and documentaries.

VAST-27M [209] consists of a total of 27 million video clips covering diverse categories, each paired with 11 captions (5 vision, 5 audio, and 1 omni-modality). The average lengths of vision, audio, and omni-modality captions are 12.5, 7.2, and 32.4 words respectively. The dataset bridges various modalities including vision, audio, and subtitles in videos. The clips were selected from the HD-VILA-100M dataset [201], ensuring each clip is between 5 and 30 seconds long and contains all three modalities. Vision captions were generated using a model trained on corpora such as MSCOCO, VATEX, MSRVTT, and MSVD [214], while audio captions were generated using VALOR-1M and WavCaps datasets. An LLM, Vicuna-13b, was used to integrate these captions into a single omni-modality caption. VAST-27M spans over 15 categories, including music, gaming, education, entertainment, and animals. its comprehensiveness, the dataset may inherit biases from the corpora and models used in its creation, highlighting the need for more diverse and larger-scale omni-modality corpora.

VI. Challenges and Future Trends

Throughout this state-of-the-art review we have analysed the most recent approaches and methodologies for the generation and detection of synthetic video and image samples. This has given us a global view of the area, as well as a glimpse of current research trends and the challenges researchers will have to face in the coming years, see Fig. 11.

First of all, we will focus on analysing the trends that will drive research in the area in the coming years, based on the results obtained from this analysis.

1. **Sample generation with diffusion models**, where the diffusion process in these models involves iterating over the input data and gradually refining the generation to fit a target distribution or to achieve the desired effect. As we have been able to observe throughout the different sections related to the generation of samples, whether video or image, the diffusion models seem to be predominating over the rest of the generation techniques, such as autoencoders or GANs. Taking into account all the research being carried out in this domain, it would not be surprising to see it monopolises multimedia content generation techniques in the coming years.
2. **Zero-Shot Learning**. This learning approach is a game changer, as it allows generative models to create content in new domains, even with entirely new features, without needing to be trained with data from those exact situations. This makes it possible, within generative techniques, to generate a wide range of content, even when a large amount of labelled data is not available. But it remains difficult to develop models capable of accurately understanding and generating content in completely new contexts. Regarding detection, zero-shot learning has the potential to help identify AI-generated content in many different data types and formats, even in the absence of huge curated datasets. However, the wide variety of synthetic content creation methods makes it difficult to create perfectly adapted detection models. Further research is needed to determine how to improve the generalisability of these models.
3. **Interpretability and Transparency**. As the content generated by AI becomes more sophisticated, it becomes increasingly important to ensure that detection models are not only effective, but also easy to understand. Users need to be convinced that the model is making the right decisions, which means that the model needs to provide clear and understandable reasons for why it has identified something as synthetic. In addition, these techniques allow us to understand whether the features that the models are using to achieve at the output are adequate or whether the system has deficiencies or biases. Therefore, the application of explainability techniques has many advantages.
4. **Multimodal data generation**. As we have seen in Section V, multimodal sample generation techniques are the least explored of all. The main reason may be their complexity, as a very precise synchronisation between video and audio has to be achieved. However, it is quite possible that this approach will start to become more relevant, due to the opportunities it presents. Regarding synthetic multimodal data detection techniques, research will be extremely limited until quality datasets are available to train robust models, capable of being applied to real situations.
5. **Model robustness**. Detection models must be able to robustly withstand various transformations and adversarial attacks, such as image compression, blurring or text paraphrasing, which can significantly degrade detection performance. The ability to withstand such manipulations is crucial for the reliable identification of synthetic content in various real-world scenarios. These types of distortions can effectively compromise a model's ability to correctly identify synthetic content. So being able to overcome these challenges is essential to ensure that the model works reliably in all kinds of scenarios.

Finally, we are going explore the different ***challenges*** that the field of video and image generation is likely to face. This review has highlighted several weaknesses that must be addressed, as they represents significant obstacles for future research in this domain.

1. **Temporal Consistency**. One of the main problems in the generation of synthetic video samples is the formation of artefacts or inconsistencies between the created frames. Smooth and realistic motion patterns are essential for video sequences, however generative models may find it difficult to maintain this from frame to frame. In addition, inconsistent frame transitions can lead to visual artifacts such as flicker, which affect the realism of the generated content. Although advances in techniques such as Implicit Neural Representations (INR), interplacing of multiple temporal attention layers, fully fine tuning on video datasets, as well as hierarchical discriminators have shown promise, further research is necessary to achieve smooth and realistic video sequences.
2. **Computational Requirements**. Video generation and detection involves processing high dimensional data, which significantly increases the computational requirements for training and inference, which can be an obstacle for small organizations. Developing more efficient algorithms and parallelization techniques for video generation is an ongoing challenge.
3. **Constant adaptation**: as we have seen in this survey, there are two main lines of research: the generation of synthetic samples and their detection techniques. Every day there are new, more sophisticated generation techniques that generate more realistic samples, so new detection models that are capable of distinguishing these synthetic samples from the real ones have to be constantly developed, i.e. it is a race. As well as the development of new quality datasets that will be the starting point of the detection systems. Another approach may be the periodic retraining of models. Whether to simply re-train a model from scratch or continue to update it through continuous learning is an ongoing challenge that researchers are still working on.
4. **Generalizability of Detection Models.** A key challenge for detection models is to be able to handle new data and new models. Generative AI models (GAIMs) evolve rapidly and if a detection model is too focused on the specific data it has been trained on, it tends to struggle with new, unseen data and updated models. To remain relevant and effective, detection models must be able to generalise to different datasets and types of generative architectures.
5. **Ethical Aspects**. The realistic nature of AI-generated content raises serious ethical questions, particularly when it comes to potential misuse. Deepfakes, fake news and other misleading content can cause real harm. To combat this, it is not enough to develop effective detection methods. We also need ethical guidelines, regulations and access controls to prevent AI technology from being used in harmful ways.

VII. Conclusions

Generative AI has witnessed exponential growth in recent years, exemplified by tools like ChatGPT that showcase its advancing capabilities. Multimedia content generation models have achieved remarkable performance across a variety of tasks, offering substantial benefits to domains such as entertainment, education, and cybersecurity. However, these advancements also introduce risks that cannot be ignored. Alongside the development of new generative AI models for producing high-quality multimedia content, there is a critical need to create detection systems that can be effectively applied in real-world situations.

This review aims to address these dual objectives by providing a comprehensive analysis of synthetic image and video generation techniques, as well as the methods used for their detection. It also

examines the principal datasets available in the current state of the art and explores future trends and challenges faced by researchers in the field. By critically evaluating the existing technologies for generating and detecting multimedia content, we seek to define the research directions that should be pursued in the coming years. The insights gathered from this survey are intended to facilitate and stimulate further research on generative AI techniques for multimedia content, ultimately contributing to both the advancement of the field and the mitigation of associated risks.

Acknowledgment

This work has been partially supported by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program; by MCIN/AEI/10.13039/501100011033/ and European Union NextGenerationEU/PRTR for XAI-Disinfodemics (PLEC 2021-007681) grant, by European Comission under IBERIFIER Plus - Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY- 04-EDMO-HUBS 101158511), and by TUAI Project (HORIZON-MSCA-2023-DN-01-01, Proposal number: 101168344); by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI; and by Comunidad Autonoma de Madrid, CIRMA-CM Project (TEC-2024/COM-404). Abdenour Hadid is funded by TotalEnergies collaboration agreement with Sorbonne University Abu Dhabi.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[2] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10741.

[3] Midjourney, "Midjourney platform." Online. [Online]. Available: https://www.midjourney.com/home, Accessed: Nov. 07, 2024.

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.

[5] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung, H. Adam, H. Akbari, Y. Alon, V. Birodkar, *et al.*, "Videopoet: A large language model for zero-shot video generation," *arXiv preprint arXiv:2312.14125*, 2023.

[6] OpenAI, "Sora: Video generation models as world simulators," OpenAI, 2024. [Online]. Available: https://openai.com/index/sora/, Accessed: Nov. 07, 2024.

[7] J. Bruce, M. D. Dennis, A. Edwards, J. Parker- Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, *et al.*, "Genie: Generative interactive environments," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*, 21–27 Jul 2024, pp. 4603–4623, PMLR.

[8] G. Madaan, S. K. Asthana, J. Kaur, "Generative ai: Applications, models, challenges, opportunities, and future directions," *Generative AI and Implications for Ethics, Security, and Data Management*, pp. 88–121, 2024.

[9] X. Zhao, X. Zhao, "Application of generative artificial intelligence in film image production," *Computer- Aided Design & Applications*, vol. 21, pp. 29–43, 2024, doi: 10.14733/cadaps.2024.S27.29-43.

[10] Á. Huertas-García, H. Liz, G. Villar-Rodríguez, Martín, J. Huertas-Tato, D. Camacho, "Aida- upm at semeval-2022 task 5: Exploring multimodal late information fusion for multimedia automatic misogyny identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 771–779.

[11] N. Anantrasirichai, D. Bull, "Artificial intelligence in the creative industries: a review," *Artificial intelligence review*, vol. 55, no. 1, pp. 589–656, 2022.

[12] H. Choi, *Generative AI Art Exploration and Image Generation Fine Tuning Techniques*. PhD dissertation, California Institute of the Arts.

[13] A. Doe, B. Smith, C. White, "Gans for medical image synthesis: A comprehensive review," *Medical Image Analysis*, vol. 78, p. 102345, 2023.

[14] U. Mittal, S. Sai, V. Chamola, *et al.*, "A comprehensive review on generative ai for education," *IEEE Access*, vol. 12, pp. 142733–142759, 2024.

[15] H. S. Mavikumbure, V. Cobilean, C. S. Wickramasinghe, D. Drake, M. Manic, "Generative ai in cyber security of cyber physical systems: Benefits and threats," in *2024 16th International Conference on Human System Interaction (HSI)*, 2024, pp. 1–8, IEEE.

[16] S. Oh, T. Shon, "Cybersecurity issues in generative ai," in *2023 International Conference on Platform Technology and Service (PlatCon)*, 2023, pp. 97–100, IEEE.

[17] H. Liz-Lopez, M. Keita, A. Taleb-Ahmed, A. Hadid, J. Huertas-Tato, D. Camacho, "Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges," *Information Fusion*, vol. 103, p. 102103, 2024.

[18] A. Giron, J. Huertas-Tato, D. Camacho, "Multimodal analysis for identifying misinformation in social networks," in *The 2024 World Congress on Information Technology Applications and Services*, 2024, World IT Congress 2024.

[19] K. Shiohara, T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18720–18729.

[20] A. Martín, A. Hernández, M. Alazab, J. Jung, D. Camacho, "Evolving generative adversarial networks to improve image steganography," *Expert Systems with Applications*, vol. 222, p. 119841, 2023.

[21] Á. Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, "Camouflage is all you need: Evaluating and enhancing transformer models robustness against camouflage adversarial attacks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[22] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6259–6276, 2022.

[23] S. Tyagi, D. Yadav, "A detailed analysis of image and video forgery detection techniques," *The Visual Computer*, vol. 39, no. 3, pp. 813–833, 2023.

[24] Z. Jia, Z. Zhang, L. Wang, T. Tan, "Human image generation: A comprehensive survey," *ACM Computing Surveys*, 2022.

[25] A. Figueira, B. Vaz, "Survey on synthetic data generation, evaluation methods and gans," *Mathematics*, vol. 10, no. 15, p. 2733, 2022.

[26] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.

[27] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. Foster, "Comprehensive exploration of synthetic data generation: A survey," *arXiv preprint arXiv:2401.02524*, 2024.

[28] P. Cao, F. Zhou, Q. Song, L. Yang, "Controllable generation with text-to-image diffusion models: A survey," *arXiv preprint arXiv:2403.04279*, 2024.

[29] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, A. Dantcheva, "Synthetic data in human analysis: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 4957–4976, 2024, doi: 10.1109/TPAMI.2024.3362821.

[30] P. Cao, F. Zhou, Q. Song, L. Yang, "Controllable generation with text-to-image diffusion models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2403.04279.

[31] T. Zhang, Z. Wang, J. Huang, M. M. Tasnim, W. Shi, "A survey of diffusion based image generation models: Issues and their solutions," 2023. [Online]. Available: https://arxiv.org/abs/2308.13142.

[32] A. Sauer, T. Karras, S. Laine, A. Geiger, T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," in *International conference on machine learning*, 2023, pp. 30105–30118, PMLR.

[33] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, T. Park, "Scaling up gans for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10124–10134.

[34] H. Ku, M. Lee, "Textcontrolgan: Text-to-image synthesis with controllable generative adversarial networks," *Applied Sciences*, vol. 13, no. 8, p. 5098, 2023.

[35] M. Tao, B.-K. Bao, H. Tang, C. Xu, "Galip: Generative adversarial clips for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, 2023, pp. 14214–14223.

[36] Y. A. Ahmed, A. Mittal, "Unsupervised co-generation of foreground-background segmentation from text- to-image synthesis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, vol. 12, 2024, pp. 5058–5069.

[37] Y. Xu, Y. Zhao, Z. Xiao, T. Hou, "Ufogen: You forward once large scale text-to-image generation via diffusion gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 8196–8206.

[38] J. Shi, C. Wu, J. Liang, X. Liu, N. Duan, "Divae: Photorealistic images synthesis with denoising diffusion decoder," *arXiv preprint arXiv:2206.00386*, 2022.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763, PMLR.

[40] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, D. Krishnan, "Muse: Text-to-image generation via masked generative transformers," 2023. [Online]. Available: https://arxiv.org/abs/2301.00704.

[41] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in neural information processing systems*, vol. 34, pp. 19822–19835, 2021.

[42] M. Ding, W. Zheng, W. Hong, J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16890– 16902, 2022.

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[44] A. Razzhigaev, A. Shakhmatov, A. Maltseva, Arkhipkin, I. Pavlov, I. Ryabov, A. Kuts, Panchenko, A. Kuznetsov, D. Dimitrov, "Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion," *arXiv preprint arXiv:2310.03502*, 2023.

[45] J. Yang, J. Feng, H. Huang, "Emogen: Emotional image content generation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6358–6368.

[46] H. Li, C. Shen, P. Torr, V. Tresp, J. Gu, "Self- discovering interpretable diffusion latent directions for responsible text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12006–12016.

[47] J. Ho, T. Salimans, "Classifier-free diffusion guidance," 2022. [Online]. Available: https://arxiv.org/abs/2207.12598.

[48] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[49] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, P. Luo, "Raphael: Text-to-image generation via large mixture of diffusion paths," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[50] G. DeepMind, "Imagen 2." http://tinyurl.com/3pakj3mk, 2023.

[51] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn.openai.com/papers/dall-e- 3.pdf*, vol. 2, no. 3, p. 8, 2023.

[52] L. Chen, W. Zhao, L. Xu, "Augmented cyclegan for enhanced image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2345–2354.

[53] Y. Wang, K. Liu, H. Zhang, "Dualgan++: Robust and efficient image-to-image translation," *IEEE Transactions on Image Processing*, vol. 32, pp. 678–690, 2023.

[54] M. Li, E. Johnson, R. Wang, "Cut++: Enhanced contrastive unpaired translation for image synthesis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 3456–3465.

[55] T. Nguyen, W. Huang, S. Lee, "Spade++: Spatially- adaptive gans for high-resolution image synthesis," *Pattern Recognition*, vol. 122, pp. 108–119, 2022.

[56] S. Kim, D. Park, M. Lee, "Self-supervised image translation gan for high-quality synthetic image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4567–4576.

[57] H. Zhang, Y. Wang, K. Liu, "Unified multimodal gan for diverse image-to-image translation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 234–245, 2024.

[58] M. Lee, S. Kim, D. Park, "Zero-shot gans: Generating images without extensive labeled data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 567–578, 2024.

[59] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, "Alias-free generative adversarial networks," in *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[60] T. Karras, T. Aila, S. Laine, J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[61] J. Smith, J. Doe, A. Brown, "Efficientgan: Reducing the computational cost of gans while preserving image quality," *Journal of Machine Learning Research*, vol. 23, pp. 1234–1256, 2022.

[62] E. Johnson, R. Wang, M. Li, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1204–1213.

[63] D. Torbunov, Y. Huang, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, Y. Ren, "Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image- to-image translation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 702–712.

[64] W. Harvey, S. Naderiparizi, F. Wood, "Conditional image generation by conditioning variational auto- encoders," *arXiv preprint arXiv:2102.12037*, 2022.

[65] A. Razavi, A. van den Oord, O. Vinyals, "Hierarchical variational autoencoders for high-resolution image synthesis," *Nature*, vol. 570, pp. 234–239, 2022.

[66] A. Vahdat, J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *arXiv preprint arXiv:2007.03898*, 2022.

[67] J.-Y. Zhu, T. Park, A. A. Efros, "Stylevae: Variational autoencoders with style transfer for image synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 2345–2356, 2023.

[68] H. Kim, A. Mnih, "Factorized hierarchical variational autoencoders for disentangled representation learning," *Journal of Machine Learning Research*, vol. 24, pp. 3456–3465, 2023.

[69] D. E. Diamantis, P. Gatoula, D. K. Iakovidis, "Endovae: Generating endoscopic images with a variational autoencoder," in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2022, pp. 1–5, IEEE.

[70] R. Dos Santos, J. Aguilar, "A synthetic data generation system based on the variational-autoencoder technique and the linked data paradigm," *Progress in Artificial Intelligence*, pp. 1–15, 2024.

[71] S. An, J.-J. Jeon, "Distributional learning of variational autoencoder: Application to synthetic data generation," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 57825–57851, Curran Associates, Inc.

[72] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, J.-Y. Zhu, "Zero-shot image-to-image translation," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[73] A. Brock, J. Donahue, K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," 2019. [Online]. Available: https://arxiv.org/abs/1809.11096.

[74] S. Sinitsa, O. Fried, "Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4067–4076.

[75] N. Poredi, D. Nagothu, Y. Chen, "Ausome: authenticating social media images using frequency analysis," in *Disruptive Technologies in Information Sciences VII*, vol. 12542, 2023, pp. 44–56, SPIE.

[76] Q. Bammey, "Synthbuster: Towards detection of diffusion model generated images," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 1–9, 2023, doi: 10.1109/OJSP.2023.3337714.

[77] T. Alzantot, C. Shou, M. Farag, Z. J. Wang, S. Pandey, M. Esmaili, "Wavelet-packets for deepfake image analysis and detection," *Machine Learning*, vol. 111, no. 11, pp. 1–25, 2022, doi: 10.1007/s10994-022-06225-5.

[78] N. Zhong, Y. Xu, Z. Qian, X. Zhang, "Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection," *arXiv*

preprint arXiv:2311.12397, 2023.

[79] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22445–22455.

[80] R. Ma, J. Duan, F. Kong, X. Shi, K. Xu, "Exposing the fake: Effective diffusion-generated images detection," *arXiv preprint arXiv:2307.06272*, 2023.

[81] J. Huertas-Tato, A. Martín, J. Fierrez, D. Camacho, "Fusing cnns and statistical indicators to improve image classification," *Information Fusion*, vol. 79, pp. 174–187, 2022.

[82] P. Lorenz, R. L. Durall, J. Keuper, "Detecting images generated by deep diffusion models using their local intrinsic dimensionality," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 448–459.

[83] L. Guarnera, O. Giudice, S. Battiato, "Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models," *arXiv preprint arXiv:2303.00608*, 2023.

[84] D. A. Coccomini, A. Esuli, F. Falchi, C. Gennaro, G. Amato, "Detecting images generated by diffusers," *PeerJ Computer Science*, vol. 10, p. e2127, 2024.

[85] U. Ojha, Y. Li, Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24480–24489.

[86] M. Mathys, M. Willi, R. Meier, "Synthetic photography detection: A visual guidance for identifying synthetic images created by ai," *arXiv preprint arXiv:2408.06398*, 2024.

[87] C. Tan, R. Tao, H. Liu, G. Gu, B. Wu, Y. Zhao, Y. Wei, "C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection," *arXiv preprint arXiv:2408.09647*, 2024.

[88] M. Keita, W. Hamidouche, H. B. Eutamene, Hadid, A. Taleb-Ahmed, "Bi-lora: A vision- language approach for synthetic image detection," *Pattern Recognition*, 2024. Preprint available at https://github.com/Mamadou-Keita/VLM-DETECT.

[89] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[90] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[91] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make- a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[92] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *arXiv preprint arXiv:2211.11018*, 2022.

[93] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, *et al.*, "Language model beats diffusion– tokenizer is key to visual generation," *arXiv preprint arXiv:2310.05737*, 2023.

[94] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[95] J. Huertas-Tato, A. Martín, D. Camacho, "Understanding writing style in social media with a supervised contrastively pre-trained transformer," *Knowledge-Based Systems*, vol. 296, p. 111867, 2024.

[96] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, I. Misra, "Emu video: Factorizing text-to-video generation by explicit image conditioning," *arXiv preprint arXiv:2311.10709*, 2023.

[97] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, *et al.*, "Lavie: High-quality video generation with cascaded latent diffusion models," *arXiv preprint arXiv:2309.15103*, 2023.

[98] W. Menapace, A. Siarohin, I. Skorokhodov, E. Deyneka, T.-S. Chen, A. Kag, Y. Fang, A. Stoliar, E. Ricci, J. Ren, *et al.*, "Snap video: Scaled spatiotemporal transformers for text-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7038–7048.

[99] T. Karras, M. Aittala, T. Aila, S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in neural information processing systems*, vol. 35, pp. 26565–26577, 2022.

[100] T. Chen, L. Li, "Fit: Far-reaching interleaved transformers," *arXiv preprint arXiv:2305.12689*, 2023.

[101] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320.

[102] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li, C. Chen, Y. Qiao, "Latte: Latent diffusion transformer for video generation," *arXiv preprint arXiv:2401.03048*, 2024.

[103] X. Li, W. Chu, Y. Wu, W. Yuan, F. Liu, Q. Zhang, F. Li, H. Feng, E. Ding, J. Wang, "Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation," *arXiv preprint arXiv:2309.00398*, 2023.

[104] C. Wu, L. Huang, Q. Zhang, B. Li, L. Ji, F. Yang, G. Sapiro, N. Duan, "Godiva: Generating open-domain videos from natural descriptions," *arXiv preprint arXiv:2104.14806*, 2021.

[105] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2630–2640.

[106] W. Hong, M. Ding, W. Zheng, X. Liu, J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[107] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, N. Duan, "Nüwa: Visual synthesis pre-training for neural visual world creation," in *European conference on computer vision*, 2022, pp. 720–736, Springer.

[108] C. Wu, J. Liang, X. Hu, Z. Gan, J. Wang, L. Wang, Z. Liu, Y. Fang, N. Duan, "Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis," *arXiv preprint arXiv:2207.09814*, 2022.

[109] W. Yan, Y. Zhang, P. Abbeel, A. Srinivas, "Videogpt: Video generation using vq-vae and transformers," *arXiv preprint arXiv:2104.10157*, 2021.

[110] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22563–22575.

[111] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, *et al.*, "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[112] Y. He, T. Yang, Y. Zhang, Y. Shan, Q. Chen, "Latent video diffusion models for high-fidelity long video generation," *arXiv preprint arXiv:2211.13221*, 2022.

[113] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[114] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, L. Fei- Fei, I. Essa, L. Jiang, J. Lezama, "Photorealistic video generation with diffusion models," *arXiv preprint arXiv:2312.06662*, 2023.

[115] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, D. Erhan, "Phenaki: Variable length video generation from open domain textual descriptions," in *International Conference on Learning Representations*, 2022.

[116] Z. Xing, Q. Dai, H. Hu, Z. Wu, Y.-G. Jiang, "Simda: Simple diffusion adapter for efficient video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7827–7839.

[117] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao, M. Z. Shou, "Show-1: Marrying pixel and latent diffusion models for text-to-video generation," *arXiv preprint arXiv:2309.15818*, 2023.

[118] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15954–15964.

[119] W. Weng, R. Feng, Y. Wang, Q. Dai, C. Wang, D. Yin, Z. Zhao, K. Qiu, J. Bao, Y. Yuan, *et al.*, "Art-v: Auto-regressive text-to-video generation with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7395–7405.

[120] F. Shi, J. Gu, H. Xu, S. Xu, W. Zhang, L. Wang, "Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and

video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024,pp. 7393–7402.

[121] Z. Qing, S. Zhang, J. Wang, X. Wang, Y. Wei, Y. Zhang, C. Gao, N. Sang, "Hierarchical spatio- temporal decoupling for text-to-video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6635–6645.

[122] R. Wu, L. Chen, T. Yang, C. Guo, C. Li, X. Zhang, "Lamp: Learn a motion pattern for few-shot-based video generation," *arXiv preprint arXiv:2310.10769*, 2023.

[123] X. Guo, M. Zheng, L. Hou, Y. Gao, Y. Deng, C. Ma, W. Hu, Z. Zha, H. Huang, P. Wan, *et al.*, "I2v-adapter: A general image-to-video adapter for video diffusion models," *arXiv preprint arXiv:2312.16693*, 2023.

[124] D. J. Zhang, D. Li, H. Le, M. Z. Shou, C. Xiong, D. Sahoo, "Moonshot: Towards controllable video generation and editing with multimodal conditions," *arXiv preprint arXiv:2401.01827*, 2024.

[125] L. Gong, Y. Zhu, W. Li, X. Kang, B. Wang, T. Ge, B. Zheng, "Atomovideo: High fidelity image-to-video generation," *arXiv preprint arXiv:2403.01800*, 2024.

[126] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin, *et al.*, "Motion-i2v: Consistent and controllable image-to- video generation with explicit motion modeling," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[127] W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, S. Huang, W. Chen, "Consisti2v: Enhancing visual consistency for image-to-video generation," *arXiv preprint arXiv:2402.04324*, 2024.

[128] C. Shen, Y. Gan, C. Chen, X. Zhu, L. Cheng, T. Gao, J. Wang, "Decouple content and motion for conditional image-to-video generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 4757–4765.

[129] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animationmagic," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.

[130] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490.

[131] M. Dorkenwald, T. Milbich, A. Blattmann, R. Rombach, K. G. Derpanis, B. Ommer, "Stochastic image-to-video synthesis using cinns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3742–3753.

[132] H. Ni, C. Shi, K. Li, S. X. Huang, M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18444–18455.

[133] C. Wang, J. Gu, P. Hu, S. Xu, H. Xu, X. Liang, "Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance," *arXiv preprint arXiv:2312.03018*, 2023.

[134] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, J. Zhou, "I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models," *arXiv preprint arXiv:2311.04145*, 2023.

[135] A. Blattmann, T. Milbich, M. Dorkenwald, B. Ommer, "Understanding object dynamics for interactive image-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5171–5181.

[136] W. Menapace, S. Lathuiliere, S. Tulyakov, A. Siarohin, E. Ricci, "Playable video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10061–10070.

[137] H. Wang, M. Huang, D. Wu, Y. Li, W. Zhang, "Supervised video-to-video synthesis for single human pose transfer," *IEEE Access*, vol. 9, pp. 17544–17556, 2021.

[138] L. Zhuo, G. Wang, S. Li, W. Wu, Z. Liu, "Fast- vid2vid: Spatial-temporal compression for video-to- video synthesis," in *European Conference on Computer Vision*, 2022, pp. 289–305, Springer.

[139] S. Yang, Y. Zhou, Z. Liu, C. C. Loy, "Rerender a video: Zero-shot text-guided video-to-video translation," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–11.

[140] W. Wang, Y. Jiang, K. Xie, Z. Liu, H. Chen, Y. Cao, X. Wang, C. Shen, "Zero-shot video editing using off-the-shelf image diffusion models," *arXiv preprint arXiv:2303.17599*, 2023.

[141] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, Q. Chen, "Fatezero: Fusing attentions for zero- shot text-based video editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15932–15942.

[142] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023.

[143] Z. Hu, D. Xu, "Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet," *arXiv preprint arXiv:2307.14073*, 2023.

[144] F. Liang, B. Wu, J. Wang, L. Yu, K. Li, Y. Zhao, I. Misra, J.-B. Huang, P. Zhang, P. Vajda, *et al.*, "Flowvid: Taming imperfect optical flows for consistent video- to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8207–8216.

[145] B. Wu, C.-Y. Chuang, X. Wang, Y. Jia, K. Krishnakumar, T. Xiao, F. Liang, L. Yu, P. Vajda, "Fairy: Fast parallelized instruction-guided video- to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8261–8270.

[146] M. Ku, C. Wei, W. Ren, H. Yang, W. Chen, "Anyv2v: A plug-and-play framework for any video-to-video editing tasks," *arXiv preprint arXiv:2403.14468*, 2024.

[147] W. Ouyang, Y. Dong, L. Yang, J. Si, X. Pan, "I2vedit: First-frame-guided video editing via image-to-video diffusion models," *arXiv preprint arXiv:2405.16537*, 2024.

[148] H. Ouyang, Q. Wang, Y. Xiao, Q. Bai, J. Zhang, K. Zheng, X. Zhou, Q. Chen, Y. Shen, "Codef: Content deformation fields for temporally consistent video processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8089–8099.

[149] Y. Gu, Y. Zhou, B. Wu, L. Yu, J.-W. Liu, R. Zhao, J. Z. Wu, D. J. Zhang, M. Z. Shou, K. Tang, "Videoswap: Customized video subject swapping with interactive semantic point correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7621–7630.

[150] J. Bai, T. He, Y. Wang, J. Guo, H. Hu, Z. Liu, J. Bian, "Uniedit: A unified tuning-free framework for video motion and appearance editing," *arXiv preprint arXiv:2402.13185*, 2024.

[151] Y. Hu, C. Luo, Z. Chen, "Make it move: controllable image-to-video generation with text descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18219–18228.

[152] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[153] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang, *et al.*, "Nuwa-xl: Diffusion over diffusion for extremely long video generation," *arXiv preprint arXiv:2303.12346*, 2023.

[154] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, Germanidis, "Structure and content-guided video synthesis with diffusion modelss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7346–7356.

[155] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, N. Duan, "Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory," *arXiv preprint arXiv:2308.08089*, 2023.

[156] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[157] H. Ni, B. Egger, S. Lohit, A. Cherian, Y. Wang, T. Koike-Akino, S. X. Huang, T. K. Marks, "Ti2v- zero: Zero-shot image conditioning for text-to-video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9015–9025.

[158] C. Nash, J. Carreira, J. Walker, I. Barr, A. Jaegle, M. Malinowski, P. Battaglia, "Transframer: Arbitrary frame prediction with generative models," *arXiv preprint arXiv:2203.09494*, 2022.

[159] D. S. Vahdati, T. D. Nguyen, A. Azizpour, M. C. Stamm, "Beyond deepfake images: Detecting ai- generated videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4397–4408.

[160] P. He, L. Zhu, J. Li, S. Wang, H. Li, "Exposing ai- generated videos: A

benchmark dataset and a local- and-global temporal defect based detection method," *arXiv preprint arXiv:2405.04133*, 2024.

[161] Z. Peng, L. Dong, H. Bao, Q. Ye, F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022.

[162] H. Chen, Y. Hong, Z. Huang, Z. Xu, Z. Gu, Y. Li, J. Lan, H. Zhu, J. Zhang, W. Wang, *et al.*, "Demamba: Ai- generated video detection on million-scale genvideo benchmark," *arXiv preprint arXiv:2405.19707*, 2024.

[163] J. Bai, M. Lin, G. Cao, "Ai-generated video detection via spatio-temporal anomaly learning," *arXiv preprint arXiv:2403.16638*, 2024.

[164] L. Ma, J. Zhang, H. Deng, N. Zhang, Y. Liao, H. Yu, "Decof: Generated video detection via frame consistency," *arXiv preprint arXiv:2402.02085*, 2024.

[165] L. Ji, Y. Lin, Z. Huang, Y. Han, X. Xu, J. Wu, C. Wang, Z. Liu, "Distinguish any fake videos: Unleashing the power of large-scale data and motion features," *arXiv preprint arXiv:2405.15343*, 2024.

[166] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8121–8130.

[167] Q. Liu, P. Shi, Y.-Y. Tsai, C. Mao, J. Yang, "Turns out i'm not real: Towards robust detection of ai-generated videos," *arXiv preprint arXiv:2406.09601*, 2024.

[168] J. Ricker, S. Damm, T. Holz, A. Fischer, "Towards the detection of diffusion model deepfakes," *arXiv preprint arXiv:2210.14571*, 2022.

[169] H. Song, S. Huang, Y. Dong, W.-W. Tu, "Robustness and generalizability of deepfake detection: A study with diffusion models," *arXiv preprint arXiv:2309.02218*, 2023.

[170] L. Papa, L. Faiella, L. Corvitto, L. Maiano, I. Amerini, "On the use of stable diffusion for creating realistic faces: From generation to detection," in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, 2023, pp. 1–6, IEEE.

[171] Y. Wang, Z. Huang, X. Hong, "Benchmarking deepart detection," *arXiv preprint arXiv:2302.14475*, 2023.

[172] Z. Sha, Z. Li, N. Yu, Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 3418–3432.

[173] Z. Xi, W. Huang, K. Wei, W. Luo, P. Zheng, "Ai- generated image detection using a cross-attention enhanced dual-stream network," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 1463– 1470, IEEE.

[174] M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, S. A. Fattah, "Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2200–2204, IEEE.

[175] S. Jia, M. Huang, Z. Zhou, Y. Ju, J. Cai, S. Lyu, "Autosplice: A text-prompt manipulated image dataset for media forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 893–903.

[176] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, X. Liu, "Hierarchical fine-grained image forgery detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3155–3165.

[177] G. Zingarini, D. Cozzolino, R. Corvi, G. Poggi, L. Verdoliva, "M3dsynth: A dataset of medical 3d images with ai-generated local manipulations," *arXiv preprint arXiv:2309.07973*, 2023.

[178] R. Shao, T. Wu, Z. Liu, "Detecting and grounding multi-modal media manipulation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6904–6913.

[179] R. Amoroso, D. Morelli, M. Cornia, L. Baraldi, A. Del Bimbo, R. Cucchiara, "Parents and children: Distinguishing multimodal deepfakes from natural images," *arXiv preprint arXiv:2304.00500*, 2023.

[180] H. Cheng, Y. Guo, T. Wang, L. Nie, M. Kankanhalli, "Diffusion facial forgery detection," *arXiv preprint arXiv:2401.15859*, 2024.

[181] J. J. Bird, A. Lotfi, "Cifake: Image classification and explainable identification of ai-generated synthetic images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024, doi: 10.1109/ACCESS.2024.3356122.

[182] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, Y. Wang, "Genimage: A million- scale benchmark for detecting ai-generated image," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[183] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, W. Ouyang, "Seeing is not always believing: benchmarking human and model perception of ai-generated images," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[184] Y. Hong, J. Zhang, "Wildfake: A large-scale challenging dataset for ai-generated images detection," *arXiv preprint arXiv:2402.11843*, 2024.

[185] S. Changpinyo, P. Sharma, N. Ding, R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre- training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.

[186] P. Sharma, N. Ding, S. Goodman, R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[187] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 2443–2449.

[188] K. Desai, G. Kaul, Z. Aysola, J. Johnson, "Redcaps: Web-curated image-text data created by the people, for the people," *arXiv preprint arXiv:2111.11431*, 2021.

[189] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image- text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022.

[190] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, Hoover, D. H. Chau, "Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models," *arXiv preprint arXiv:2210.14896*, 2022.

[191] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[192] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, "Raise: A raw images dataset for digital image forensics," in *Proceedings of the 6th ACM multimedia systems conference*, 2015, pp. 219–224.

[193] Clip-interrogator, "Clip-interrogator," 2022. Available: https://github.com/pharmapsychotic/ clip-interrogator.

[194] ALASKA, "Alaska." https://alaska.utt.fr/. Accessed: 2024-08-04.

[195] OpenAI, "Dall·e 2." https://openai.com/product/dall-e-2. Accessed: 2024-08-04.

[196] DreamStudio, "Dreamstudio." https://beta. dreamstudio.ai/generate. Accessed: 2024-08-04.

[197] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images." https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf, 2009.

[198] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in neural information processing systems*, vol. 34, pp. 23634– 23651, 2021.

[199] M. Bain, A. Nagrani, G. Varol, A. Zisserman, "Frozen in time: A joint video and image encoder for end- to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.

[200] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, C. C. Loy, "Celebv-hq: A large-scale video facial attributes dataset," in *European conference on computer vision*, 2022, pp. 650–667, Springer.

[201] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5036–5045.

[202] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023.

[203] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, W. Wu, "Celebv-text: A large-scale facial text-video dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14805–14814.

[204] W. Wang, H. Yang, Z. Tuo, H. He, J. Zhu, J. Fu, J. Liu, "Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation," https://openreview.net/forum?id=dUDwK38MVC, 2023.

[205] H. Xu, Q. Ye, X. Wu, M. Yan, Y. Miao, J. Ye, G. Xu, A. Hu, Y. Shi, G. Xu, *et al.*, "Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks," *arXiv preprint arXiv:2306.04362*, 2023.

[206] W. Wang, Y. Yang, "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models," *arXiv preprint arXiv:2403.06098*, 2024.

[207] X. Ju, Y. Gao, Z. Zhang, Z. Yuan, X. Wang, A. Zeng, Y. Xiong, Q. Xu, Y. Shan, "Miradata: A large-scale video dataset with long durations and structured captions," *arXiv preprint arXiv:2407.06358*, 2024.

[208] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-W. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang, *et al.*, "Panda-70m: Captioning 70m videos with multiple cross-modality teachers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13320–13331.

[209] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, J. Liu, "Vast: A vision-audio-subtitle-text omni- modality foundation model and dataset," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[210] R. Girdhar, D. Ramanan, "Cater: A diagnostic dataset for compositional actions and temporal reasoning," *arXiv preprint arXiv:1910.04744*, 2019.

[211] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[212] L. Han, J. Ren, H.-Y. Lee, F. Barbieri, K. Olszewski, S. Minaee, D. Metaxas, S. Tulyakov, "Show me what and tell me how: Video synthesis via multimodal conditioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3615–3625.

[213] J. Xu, T. Mei, T. Yao, Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[214] D. Chen, W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[215] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, A. Oliva, "Spoken moments: Learning joint audio-visual representations from video descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14871–14881.

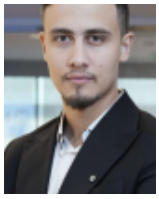
[216] B. C. Hosler, X. Zhao, O. Mayer, C. Chen, J. A. Shackleford, M. C. Stamm, "The video authentication and camera identification database: A new database for video forensics," *IEEE access*, vol. 7, pp. 76937– 76948, 2019.

[217] Moonvalley, "Moonvalley - ai video generation," 2024. [Online]. Available: https://moonvalley.ai/, Accessed: 2024-08-16.

[218] L. Yang, Y. Fan, N. Xu, "Video instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5188–5197.

[219] L. Huang, X. Zhao, K. Huang, "Got-10k: A large high- diversity benchmark for generic object tracking in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[220] X. Shang, T. Ren, J. Guo, H. Zhang, T.-S. Chua, "Video visual relation detection," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1300–1308.



Hessen Bougueffa

Hessen Bougueffa graduated with a Master's degree in Telecommunication Systems in 2022. His Master's thesis, "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," set a strong foundation in applying advanced computational techniques to real-world problems. Presently, as a Ph.D. candidate at Polytechnic Haute-de-France, Hessen is working on the development of multimodal models for content characterization in collaboration with the Martini Project. His research is carving a niche at the crossroads of machine learning and content analysis, exploring how various data types can be synergistically utilized for enhanced content understanding. Hessen's work is expected to contribute significantly to the fields of artificial intelligence and data science, pushing the envelope in multimodal learning approaches.

Mamadou Keita

Mamadou Keita received his Engineer's degree in Telecommunications and Computer Networks from the National Institute of Telecommunications and Information Technology in Oran, Algeria, the Master's degree in Engineering and Innovation in Images and Networks with a specialization in Images from Sorbonne Paris Nord University, France in 2022. He is currently pursuing a Ph.D. degree in signal processing at the Institute of Electronics, Microelectronics and Nanotechnology, Polytechnic University of Hauts de France, Valenciennes, France.His research interests include image quality assessment, object detection and tracking, object segmentation, behavior analysis, medical imaging, and multimedia security.

Wassim Hamidouche

Wassim Hamidouche is a Principal Researcher at Technology Innovation Institute (TII) in Abu Dhabi, UAE. He also holds the position of Associate Professor at INSA Rennes and is a member of the Institute of Electronics and Telecommunications of Rennes (IETR), UMR CNRS 6164. He earned his Ph.D. degree in signal and image processing from the University of Poitiers, France, in 2010. From 2011 to 2012, he worked as a Research Engineer at the Canon Research Centre in Rennes, France. Additionally, he served as a researcher at the IRT b<>com research Institute in Rennes from 2017 to 2022. He has over 180 papers published in the field of image processing and computer vision. His research interests encompass various areas, including video coding, the design of software and hardware circuits and systems for video coding standards, image quality assessment, and multimedia security.

Abdelmalik Taleb-Ahmed

Abdelmalik Taleb-Ahmed received in 1992 PhD in electronics and microelectronics from université des Sciences et Technologies de Lille 1. He was associate professor in Calais until 2004. He joined the Université Polytechnique des Hauts de France in 2004, where he is presently Full Professor. He joined the laboratory IEMN DOAE. his research focused on computer vision and artificial intelligence and machine vision. His research interests include segmentation, classification, data fusion, pattern recognition, computer vision, and machine learning, with applications in biometrics, video surveillance, autonomous driving, and medical imaging. He has (co-)authored over 225 peer-reviewed papers and (co-)supervised 20 graduate students in these areas of research. His recent research revolves mainly around: Enhanced Perception and HD Mapping in intelligent Transportation, Digitalization of Road and the Signaling, E-Health and Artificial Intelligence, pattern recognition, computer vision, and information fusion, with applications in affective computing, biometrics, medical image analysis, and video analytics and surveillance.

Helena Liz-López

Helena Liz-López is an Assistant Professor in the Department of Computer Systems Engineering at the Universidad Politécnica de Madrid (UPM) and a member of the Natural Language Processing and Deep Learning (NLP&DL) research group. She holds a degree in Biology from the Universidad Autónoma de Madrid and a master's degree in bioinformatics and computational biology from the same university. She obtained her PhD in computer sciences from the Universidad Politécnica de Madrid in 2024, receiving the distinction of "cum laude." Her research interests include Deep Learning, Machine Learning applications in ecology and medicine, and explainable AI.

Alejandro Martin

Alejandro Martin is Associate Professor at Universidad Politécnica de Madrid. His main research interests are Deep Learning, Cybersecurity, and Natural Language Processing. He has been visiting researcher at theUniversity of Kent and the University of Córdoba. Besides has participated in an important number of international conferences as a reviewer and organizer, as a reviewer and Guest Editor in international journals, and in a large number of research projects. He is the PI of different national and international projects focused on the application AI to detect and track misinformation in social networks.

David Camacho

David Camacho received the Ph.D. degree (with Honors) in Computer Xcience from Universidad Carlos III de Madrid, in 2001. He is currently a Full Professor with Computer Systems Engineering Department, Universidad Politécnica de Madrid (UPM), Madrid, Spain, and the Head of the Applied Intelligence and Data Analysis research Group, UPM. He has authored or coauthored more than 300

journals, books, and conference papers. His research interests include machine learning (clustering/deep learning), computational intelligence (evolutionary computation, swarm intelligence), social network analysis, fake news and disinformation analysis. He has participated/led more than 60 research projects (Spanish and European: H2020, DG Justice, ISFP, and Erasmus+), related to the design and application of artificial intelligence methods for data mining and optimization for problems emerging in industrial scenarios, aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, or video games among others.

Abdenour Hadid

Abdenour Hadid received his Doctor of Science in Technology degree in electrical and information engineering from the University of Oulu, Finland, in 2005. Now, he is a Professor in a Chair of excellence at Sorbonne Center for Artificial Intelligence (SCAI). His research interests include computer vision, deep learning, artificial intelligence, internet of things, autonomous driving and personalized healthcare. He has authored more than 400 papers in international conferences and journals, and served as a reviewer for many international conferences and journals. His research works have been well referenced by the research community with more 25000 citations and an H-Index of 59, according to Google Scholar. Prof. Hadid was the recipient of the prestigious “Jan Koenderink Prize” for fundamental contributions in computer vision.