A Survey of Cross-Modal Visual Content Generation

Fatemeh Nazarieh, Zhenhua Feng, *Senior Member, IEEE,* Muhammad Awais, Wenwu Wang, *Senior Member, IEEE,* Josef Kittler, *Life Member, IEEE*

***Abstract*—Cross-modal content generation has become very popular in recent years. To generate high-quality and realistic content, a variety of methods have been proposed. Among these approaches, visual content generation has attracted significant attention from academia and industry due to its vast potential in various applications. This survey provides an overview of recent advances in visual content generation conditioned on other modalities, such as text, audio, speech, and music, with a focus on their key contributions to the community. In addition, we summarize the existing publicly available datasets that can be used for training and benchmarking cross-modal visual content generation models. We provide an in-depth exploration of the datasets used for audio-to-visual content generation, filling a gap in the existing literature. Various evaluation metrics are also introduced along with the datasets. Furthermore, we discuss the challenges and limitations encountered in the area, such as modality alignment and semantic coherence. Last, we outline possible future directions for synthesizing visual content from other modalities including the exploration of new modalities, and the development of multi-task multi-modal networks. This survey serves as a resource for researchers interested in quickly gaining insights into this burgeoning field.**

***Index Terms*—Generative models, cross-modal, visual content generation.**

I. Introduction

In recent years, cross-modal content generation has gained significant attention due to the rapid development of modern deep generative networks and the growth of multimodal data. This emerging field aims to generate high-quality and realistic content across different modalities, such as text, sound, speech, image, video, and 3D point cloud, by leveraging input from another modality. As vision is one of the most important senses of the human biological cognitive system, this survey specifically focuses on visual content generation conditioned by other modalities.

Despite the existence of several surveys in cross-modal visual content generation [1]–[4], each of them only focuses on a single modality, such as text-to-vision, audio-to-vision, etc. The research community lacks a comprehensive survey that provides a big picture of the whole area. To bridge this gap, this survey aims to cover the recent advances in various cross-modal visual content generation areas, making it easier for researchers to track recent studies. More importantly, the underlying methods used in cross-modal visual content generation have significant overlap so it would be important to discuss them jointly and gain a better understanding of their similarities, differences, strengths, and weaknesses. Discussing them jointly can determine potential relations between text, audio, and other modality-guided visual content generation methods, where leveraging methods from one area can enhance the performance of another.

As we know, humans have the ability to imagine a scene using information from other senses, such as hearing, touching, tasting, smelling, etc. This cognitive process showcases the ability of the human brain to bridge the gap between other modalities and vision. This has motivated researchers to explore cross-modal visual content generation by exploring possible ways of transferring information from various domains to vision, as this will pave the way for a wide range of groundbreaking applications [5]. For instance, we can automatically translate the live-streaming video of a person from a specific language to the desired one while preserving realistic lip synchronization [6]. For another example, automatic choreographing is possible by developing music-to-dance generation methods. Such a model synthesizes virtual avatars, performing movements aligned with the melody and beats of the music [7].

There is a long history of research in visual content generation. In 1980s, Chellappa *et al.* [8] proposed a two-dimensional noncausal autoregressive model for pattern generation tasks. They demonstrated that by combining shades from surrounding areas with random noise, new patterns can be generated. Similarly, Cross *et al.* [9] investigated the use of Markov Random Fields (MRF) to represent and create textures by learning the mathematical relations between patterns. Despite their functionality, these models are not capable of generating realistic images. Wei *et al.* [10] proposed an enhanced method based on MRF for texture synthesis by including tree-structured vector quantization. They extended the capability of image generation from simple textures to complex images. Despite the improvement in the generated visual content, the image quality is still far from being realistic. One primary drawback of traditional methods is their inherent tendency to capture only local structures by exclusively learning the relationship between a pixel and its neighbors. These approaches struggle with learning global structures and patterns present in many natural images, leading to results lacking realism and containing artifacts.

Later on, various machine learning models such as Support Vector Machine (SVM) [11], K-Nearest Neighbor (KNN) [12] and Boltzmann Machines [13] have been investigated. Eslami developed an object shape generation model, based on the

F. Nazarieh and Z. Feng are with the School of Computer Science and Electronic Engineering, and the Nature Inspired Computing and Engineering (NICE) research group, University of Surrey, Guildford GU2 7XH, UK.

M. Awais, W. Wang and J. Kittler are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK.

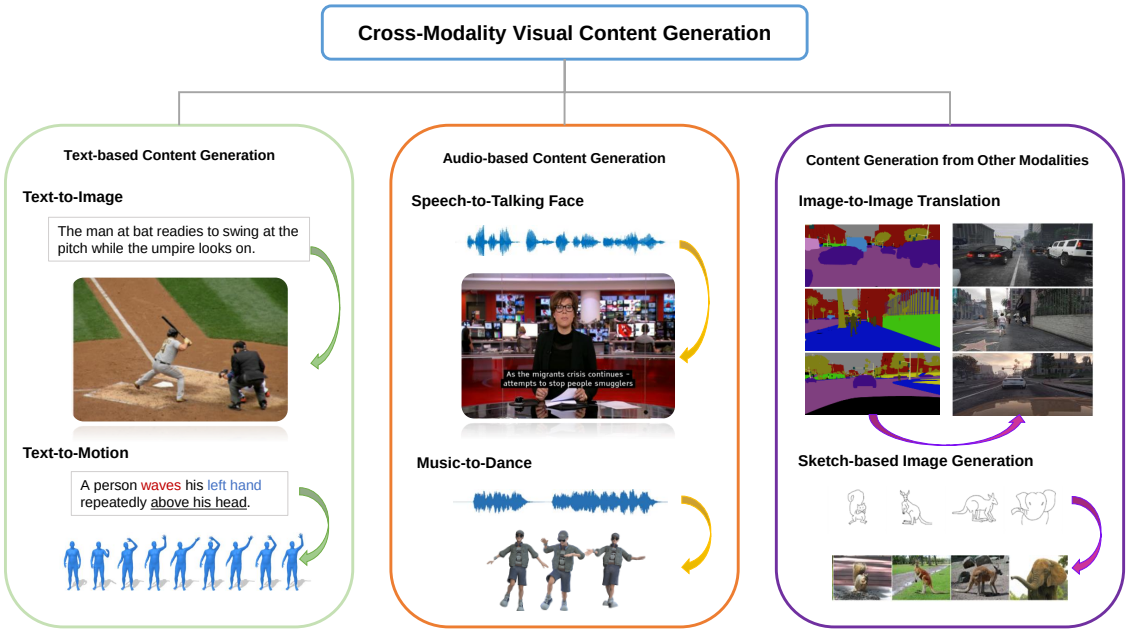


Fig. 1: The proposed taxonomy of the existing cross-modal visual content generation methods.

Boltzmann machine called the Shape Boltzmann Machine (SBM) [13]. This model learns data distribution directly from training data. Guo *et al.* proposed a Dynamic Texture Synthesis model for videos, namely Linear Dynamic Systems (LDS) [12] using Singular Value Decomposition (SVD) and KNN. However, these methods have difficulties in generating high-quality, realistic imagery and videos due to the use of handcrafted features and shallow models.

With the development of deep neural networks and their outstanding performance in various areas [14]–[16], the existing mainstream cross-modal visual content generation methods are all deep-learning-based. A deep network effectively learns the representations of visual content and captures complex patterns within the data. It also enables end-to-end learning by mapping the source data into the expected target without the need for handcrafted features as was the case with prior techniques [5], [17]. $AlignDraw$ [18] and $Speech2Vid$ [19] are the two deep-learning-based pioneering research in text-to-vision and audio-to-vision generation, respectively. $AlignDraw$ [18] uses a bidirectional Recurrent Neural Network (RNN) for textual data processing and a set of generative RNNs for image generation. In this work, an image is generated patch by patch. Using audio sequence and an identity frame, $Speech2Vid$ [19] generates talking face videos using Convolutional Neural Networks (CNNs).

With the emergence of Generative Adversarial Networks (GANs) [20], they have been widely used in different generation tasks [21], [22]. GANs present a strong ability in generating realistic and diverse samples by leveraging the adversarial training process [23]. Nevertheless, the obtained results for text-to-vision and audio-to-vision are often blurry and not realistic, especially when using newer large-scale datasets [24]–[27]. Mainly due to the lack of mechanisms in these methods to capture global coherence and intrinsic details, they fail to project information effectively from the source domain to the expected visual domain.

To address the above limitations, some researchers utilize large-scale language models (Transformers [28]) in place of traditional text encoders (e.g., RNN-based models [23]) for textual data processing [29], [30]. Assisted by the self-attention mechanism in Transformers, the relationships between words in a sentence are better captured and the generative model effectively learns the contextual dependencies and semantic structure of the input description [28], [30]. Transformers have been used for audio and visual processing as well, considerably improving the audio-to-vision generation alignment [31]. More recently, Denoising Diffusion Probabilistic Models (DDPMs) [32] have attracted widespread attention and been used in a variety of cross-modal content generation tasks [33]–[35]. As opposed to traditional generative models that rely on explicit probability distributions, diffusion models learn to generate samples by iteratively diffusing noise through a series of learnable steps and predicting the additional noises [32], [36].

Cross-modal visual content generation is at its early stage and due to the increasing number of research on this topic, keeping track of the recent state-of-the-art without a concise survey can be challenging. Also, most of the existing survey papers focus on a single modality. For instance, Agnese *et al.* provided an overview of text-to-image generation focusing on the architecture design of GAN-based models [1]. Zhang *et al.* [2] summarized recent text-to-image generation methods utilizing diffusion models, accompanied by background information on this topic and current challenges. For audio-to-talking face generation, Zhen *et al.* [3], and Tong *et al.* [4] provided an overview on generative models design, datasets, metrics and challenges.

To bridge this gap, this paper presents a comprehensive survey of state-of-the-art methods for visual content generation across different modalities. To better investigate each task, we propose a new taxonomy of the existing cross-modal visual content generation, as shown in Fig. 1. To assist researchers in identifying and understanding the challenges associated with assessing the quality and effectiveness of generated

visual content, we provide and compare current datasets and evaluation metrics. More importantly, we further discuss the existing gaps and future research directions in this field, to enhance the identification of emerging trends and push the state-of-the-art in the cross-modal visual content generation domain. To summarize, the main contributions of this survey include:

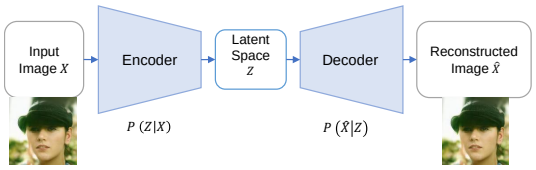
- To the best of our knowledge, this is the first overview that covers recent multimodal-guided visual content generation methods, including text-to-vision, audio-to-vision and other-modality-guided visual content generation.
- We comprehensively overview the benchmarking datasets and evaluation metrics. We also highlight the challenges associated with the existing datasets and metrics in cross-modal visual content generation.
- Open challenges and possible future directions for multimodal-guided visual content generation are identified for future work.

The remainder of this survey is organized as follows. We first provide an overview of the methods used in cross-modal visual content generation in Section II. We further review and compare the existing text-to-vision, audio-to-vision and visual content generation methods based on other modalities in Section III, Section IV, and Section V, respectively. The widely used datasets and evaluation metrics are presented in Section VI. In Section VII, we discuss the challenges in the area and propose several possible directions for further development. Last, the conclusion is drawn in Section VIII.

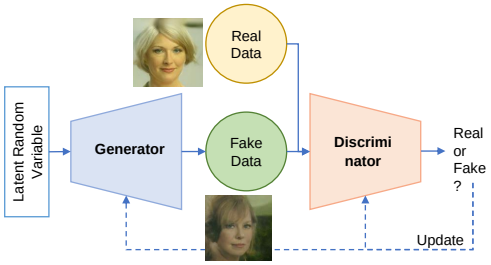
II. An Overview of The Underlying Methods Used in Cross-Modal Visual Content Generation

Although the inputs of different cross-modal generation tasks come from different domains, the methods are closely related and follow similar underlying principles. In this section, we briefly review multimodal generative models, while providing information on the most popular methods in visual content generation. The existing generative models can be roughly divided into two main categories: unimodal and multimodal models. Unimodal models generate the output using the same modality as the input data, while multimodal models generate an output of a different modality [37]. The primary connection between these tasks is the need to identify an appropriate mapping function to transform information from a source modality to the target modality so that the generated content accurately reflects the source domain information.

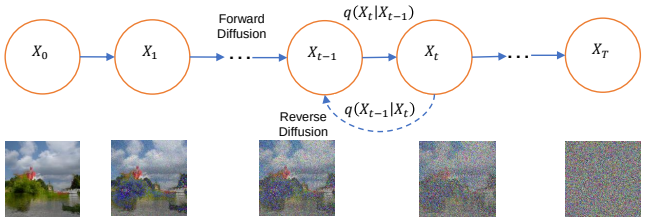
In recent years, considerable research has been conducted on generative AI, resulting in numerous generative models. As shown in Fig. 2, the three commonly employed generative frameworks include Generative Adversarial Networks (GAN), Variational AutoEncoders (VAEs), and diffusion models. GAN [20] is comprised of two main components: generator and discriminator. While the discriminator decides, whether the input comes from the real data distribution or not, the generator makes an effort to understand the distribution of the ground truth data, in order to generate realistic samples that can fool the discriminator. However, although the samples generated by GAN are of high quality, they exhibit less



(a) VAE



(b) GAN



(c) DDPM

Fig. 2: A comparison of the commonly used generative models: VAE, GAN, and DDPM.

diversity [38]. Furthermore, they suffer from unstable training, resulting in mode collapse and slow convergence [39].

VAEs are among the encoder-decoder-based generative models that learn data distribution by mapping input samples to a probabilistic distribution and then reconstructing them so that they are as close as possible to the ground truth distribution. Despite the ability of VAEs to learn the data distribution, the relatively low quality of the generated results makes them less desirable as a candidate for cross-modal visual content generation. Notably thanks to the encoder-decoder-based architecture, VAEs learn a rich latent representation. This can make them a suitable condition encoder or a prior encoder, before passing the data for processing to the generative model (e.g. VAE in stable diffusion [36]).

Denoising Diffusion Probabilistic Models (DDPMs) [32] are the new state-of-the-art for high-quality visual content generation. DDPMs are probabilistic generative models. They consist of two main phases: the forward diffusion process and the backward diffusion process. In the first stage, data is gradually corrupted into a Gaussian noise using a Markov chain process. In the reverse process, a network (such as U-Net) learns to predict the noises added at each step that should be removed in order to recover the initial image. This step is commonly known as the denoising step and a reconstruction loss is used for training DDPMs.

It should be noted that the training time of DDPM is generally longer, as compared with other models, due to

the forward and backward process. However, based on the reported results, the diffusion models can achieve excellent performance. They have advanced the state-of-the-art. One of the main reasons for delivering such a great performance is the process of learning the image structure by the steps of adding noise and subsequent denoising. This greatly assists the model in learning different parts of an image and making connections between input conditions and visual content.

Despite the recent success of diffusion models, further research development of other generative models is strongly encouraged. Nowadays there are a variety of pre-trained task-specific transformers, which can be a good starting point as pre-trained encoders. Further, earlier generative models are relatively faster in terms of the training and generation speed. Additional research is required to investigate the impact of the recent advances on earlier generative models. This will further be discussed in Section VII.

III. Text to Visual Content Generation

In this section, we first overview the mainstream methods for text-to-vision (images/videos) generation. Then we introduce the methods developed for text-to-motion generation.

A. Text-to-Image/Video

Synthesizing images and videos based on a text prompt is a popular research topic. The generated visual content is expected to reflect the text description while ensuring a realistic quality. Learning an aligned feature space for vision and language poses a great challenge for researchers since they originate from different modalities. Various studies have been conducted to address this issue. In this section, we categorize the existing methods into two subsections based on the space in which the generation process is performed.

1) Generation in the Pixel Space:

Text-to-image generation Zhang *et al.* proposed StackGAN [22] for image generation using textual descriptions. The proposed architecture has two stages. In the first stage, a conditional GAN takes the textual description as input and generates a low-resolution image that captures the overall scene and layout. To enhance the details and improve the visual quality, the second stage refines the output of the first stage conditioned on the text description. This framework allows the model to progressively refine and generate more realistic and visually coherent results. Zhang *et al.* further hypothesized that GAN training for image generation can be stabilized by breaking the generation process into sub-problems (gradually enhancing the image quality at each step). This motivated them in developing a tree-like structure based on stackGAN [22], known as StackGAN++ [40]. By incorporating conditioning augmentation techniques and multi-scale generators and discriminators, StackGAN++ achieves significant improvements in image quality and diversity compared to previous methods in text-to-image synthesis.

To improve the semantic alignment between the input description and the synthesized visual content, Xu *et al.* proposed AttnGAN [23]. This model incorporates attention mechanisms to focus on the relevant parts of the text and image during the generation process. Similar to the previous architectures, AttnGAN has two main components: text encoder and image generator. The key contribution of AttnGAN lies in the attention mechanism employed at multiple stages, allowing the generative model to access the key features extracted from the input description in order to refine the synthesized image.

(a) DALL-E1 (b) DALL-E2

(c) GLIDE classifier-free (d) GLIDE CLIP-guided (e) Stable Diffusion

Fig. 3: A comparison of popular text-to-image generation models. The input prompt is "an armchair in the shape of an avocado".

Recently, Transformers have achieved promising performance in natural language processing, such as machine translation [41] and language generation [42]. As a result, they have been widely used as text encoders in many text-to-vision generation frameworks, replacing the traditional methods. By leveraging the semantic context embedded in textual data, Transformers can be considered as valuable sources of information for fine-grained and text-aligned image generation. Naveen *et al.* addressed the challenge of generating realistic images from textual descriptions by incorporating Transformer models such as BERT, GPT2, and XLNet into AttnGAN [30]. Considering the ability of Transformers to capture semantic information and context from text, integrating them with AttnGAN enhances the text vision alignment. Ramesh *et al.* proposed a Transformer-based architecture for text-to-image generation using text and image tokens as a single stream of data. This allows the model to generate visual content from text prompts that were not used in the training stage. Therefore, it brings zero-shot learning capability into text-to-vision generation and reduces the requirement of having a large volume of accurately aligned text-image data for training. This model is commonly known as DALL-E1 [43] and it consists of two main steps. First, it uses a discrete Variational Auto Encoder (dVAE) to convert an image into multiple image tokens. In the second stage, image and text tokens are concatenated and trained in an autoregressive manner using a Transformer to learn a prior distribution over the text and image tokens. An example image generated by DALL-E1 is shown in Figure 3-(a). Although DALL-E1 generates an armchair shape, including avocado in the design, the generated object is not realistic and detail-oriented.

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [32] have demonstrated impressive results in

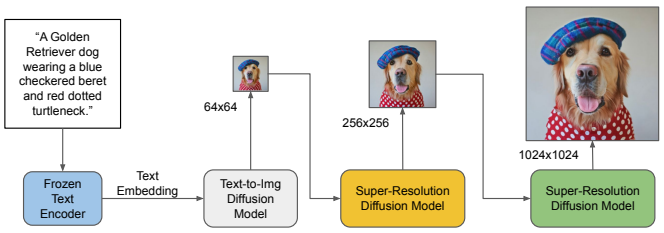


Fig. 4: The image generation pipeline of Imagen [52].

image generation. Numerous text-to-visual content generation methods [44]–[49] have employed diffusion models and obtained higher-resolution images. Guided Language to Image Diffusion for Generation and Editing (GLIDE) [50] is among the first diffusion-based text conditional image generation models. This work investigates image synthesis based on the description provided using CLIP or classifier-free approaches. CLIP guidance involves utilizing the CLIP model [51], which combines representations extracted by a vision encoder and a text encoder. The classifier-free guidance does not involve any classifier. Based on the results obtained and the feedback received from human evaluators, the synthesized results by the classifier-free guidance are more realistic and aligned with input descriptions. This is demonstrated in Fig. 3, where the image generated with the classifier-free guidance depicts an armchair that is more realistic than the one with the clip guidance. This model surpasses the state-of-the-art by the methods such as DALL-E in terms of fidelity and diversity. It should be noted that, besides text-to-image generation, GLIDE can also be used for image inpainting, providing text-guided image editing.

Imagen [52] is another text-to-image generation model that performs in the pixel space. Similar to GLIDE [50], Imagen uses a classifier-free guidance approach for visual content generation. The core difference between GLIDE and Imagen lies in their selection of text encoder and how it is used. GLIDE uses a large language model as the text encoder and trains it together with a diffusion prior and text-image pairs. In contrast, Imagen employs the frozen T5-XXL model as a text encoder to accelerate the training. Additionally, Imagen demonstrates that the use of language models, trained solely on text-only corpus, works well as a text encoder. Moreover, training a large language model on text-image pairs cannot necessarily lead to better-aligned and higher-quality images. In view of this, Imagen uses two super-resolution diffusion modules to enhance the image quality, as depicted in Fig. 4.

Text-to-Video Generation Using Transformers in video generation is often challenging due to computational costs and the scarcity of relevant text-video datasets. Hong *et al.* addressed these limitations by proposing the Transformer-based CogVideo [53], which is based on the pre-trained text-to-image CogView [29] model, enabling it to leverage the knowledge learned from the text-to-image generation domain. Also, a multi-frame-rate hierarchical strategy is used to enhance the alignment and temporal consistency between text and video content. This approach applies different frame rates at each step based on the described activities in the text, allowing the model to generate a frame sequence that fully covers the

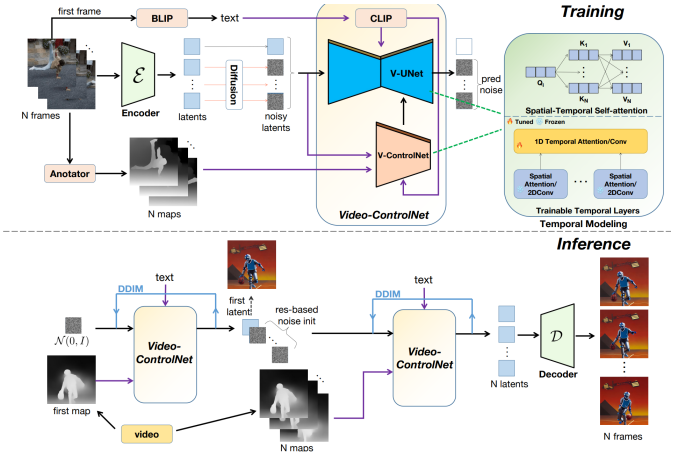


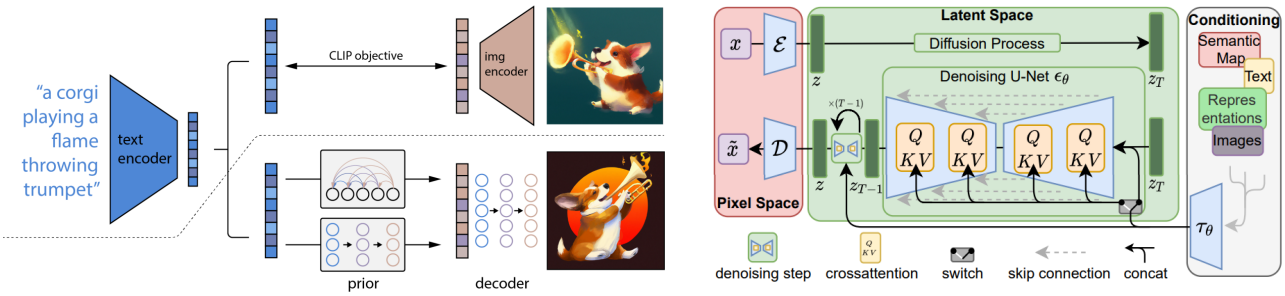
Fig. 5: The architecture of Video-ControlNet [56].

described action.

In contrast to the text-image pair data, there is a lack of large-scale text-video datasets with high-quality video frames. Make-A-Video [54] aims to leverage the progress made in text-to-image generation and apply it to the text-to-video generation domain. It decomposes the full temporal U-Net and attention tensors and approximates them in space and time. This decomposition enables effective modeling of both the spatial and temporal aspects of a video. The synthesis of high-resolution videos is enabled by a spatial-temporal pipeline that includes a video decoder, an interpolation model, and two super-resolution models.

The aforementioned studies are confined to generating fixed and short videos. To overcome this restriction, Villegas *et al.* introduced PHENAKI [55], in which a new model for learning video representations that compresses videos into discrete tokens was proposed. The extracted tokens act as video frame representations. By concatenating these tokens and de-tokenizing them, a video is created. To produce longer videos, PHENAKI conditions the frame generation process by different text descriptions at each time step. Further, to overcome the data limitations, it jointly trains on a large corpus of image-text pairs, along with a smaller number of video-text examples.

Chen *et al.* went one step further and proposed the Video-ControlNet [56], a controllable text-to-video diffusion model, using a novel strategy. As shown in Fig. 5, the method generates videos based on an input text description and uses auxiliary control signals such as edge or depth maps as guidance measures. Since the ControlNet strategy was initially proposed for text-to-image generation, a spatial-temporal attention mechanism is added to maintain temporal consistency between the generated frames. This attention mechanism enables the model to comprehend how various parts of a video relate to one another across time. Video-ControlNet drives the video sequence generation based on the initial frame and a subsequent control condition. This strategy is called first-frame conditioning. Based on the reported results, first-frame conditioning can facilitate the extension of the text-to-image generation into the text-to-video domain, enabling the production of videos of arbitrary length.



(a) DALL-E2 (b) Stable Diffusion

Fig. 6: The architectures of DALL-E2 [57] and Stable Diffusion [36].

2) Generation in Latent Space:
Although visual content generation has gained widespread popularity, synthesizing diverse, realistic and high-quality images is computationally expensive [29], [36]. To address this issue, many studies perform the generation in a latent space. Latent space works as an intermediate representation of much lower dimensionality, thus affording greater computational efficiency [36].

An improvement of DALL-E1 was proposed by using representations learned by CLIP for text-to-image generation. This model is commonly known as the DALL-E2 or unCLIP (Fig. 6). DALL-E2, leveraging contrastive models [51], consists of two main components: a "prior" and a decoder. The prior generates a CLIP image embedding given a text prompt, which is used as the image representation. The decoder generates an image conditioned on the generated image embedding in the previous step. Since the CLIP-derived representation is enriched by its multi-modal latent space, using its extracted image embedding can improve the text-vision similarity and diversity exhibited by the synthesized image. Generated image using DALL-E2 is illustrated in Fig. 3. It is evident that the generated image is of higher quality than the images generated by earlier methods and demonstrates creative design aspects. If we compare this image to the one generated by GLIDE CLIP-guided, we can see they both share a similar shape in how they portray an armchair in the form of an avocado. This might be the artifact of employing CLIP image embedding as part of image generation.

Rombach *et al.* proposed the latent diffusion model commonly known as stable diffusion [36]. Using a pre-trained variational autoencoder, the input image is transformed into a lower-dimensional representation. This latent space is then used for diffusion steps. Further, to improve the flexibility of the synthesized output, driven by the conditioning text, cross-attention layers are included in the model architecture (Fig. 6). This approach manages to trade off computational complexity for detail preservation in the synthesized visual content. As demonstrated in Fig. 3-(e), the image produced by stable diffusion is quite detailed, in addition to the ability to reflect the input prompt accurately. The generated image is of high quality and we can observe that the texture and shape of the armchair represent an avocado. VQ-diffusion [58] performs diffusion steps in the latent space extracted by VQ-VAE [59]. The diffusion process in this model follows a masked-and-replace strategy. Given a text prompt, the masking generation

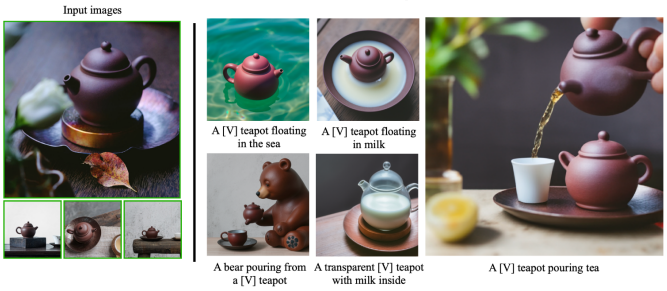


Fig. 7: A query example for DreamBooth training [34].

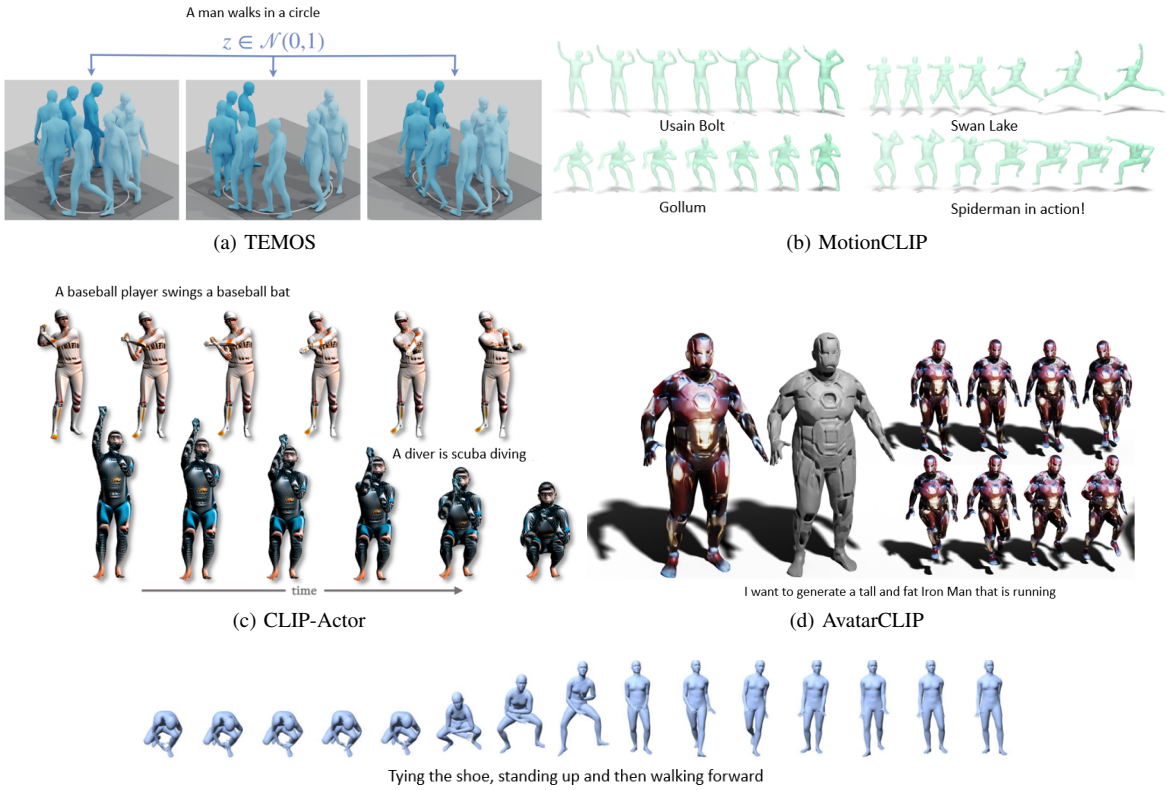
allows the model to learn better which part of the image needs to be modified.

Despite the improvement in realism and semantic alignment of generated images using the condition mechanism in stable diffusion, Chefer *et al.* [60] reported occasional neglect in generated visual content by this model. It may fail to assign a certain color to an object in the synthesized image or not include a specific object in the image (e.g., including a butterfly in the background or the existence of sunglasses). To address this limitation Chefer *et al.* introduced the Generative Semantic Nursing (GSN) method. GSN uses subject token attention maps to guide the generation process as it continuously refines the latent code over different time steps.

3) Diffusion Model and Latent Personalization:
A latent space carries considerable information for image generation. This introduces the idea of latent representation personalization in order to customize the generated image. For example, changing a desired subject attribute such as color, shape, and location, or adding new features to it.

Ruiz *et al.* proposed DreamBooth [34] that integrates the subject of interest into the model's output domain by simply fine-tuning a pre-trained text-to-image diffusion model using 3-5 images of a particular subject (unique identifier). For training this model, it is important to shape the prompt in a way that represents the user's new data, such as "a [V] teapot floating in milk" in Fig. 7. Here, [V] represents the unique identifier and "teapot" is the class name corresponding to the desired set of images used for fine-tuning. To prevent information drift, in parallel to fine-tuning on text-image pairs, a class-specific prior preservation loss is used to enforce semantic information on the class of object into the synthesized images. In this way, the model assures semantically aligned and versatile images.

Similar to DreamBooth, Gal *et al.* [61] presented a diffusion-based approach to produce user-specified concepts



(a) TEMOS (b) MotionCLIP

(c) CLIP-Actor (d) AvatarCLIP

(e) MotionDiffuse

Fig. 8: A comparison of different motion generation methods, including TEMOS [62], MotionCLIP [63], CLIP-Actor [64], AvatarCLIP [65], and MotionDiffuse [35].

that faithfully replicate the essence of the text prompts. The Text Inversion method is used to achieve this. First, text embeddings are extracted using a pre-trained text encoder, such as BERT. In this step, an empty vector is introduced to the text embedding space and is learned as part of the text encoding process to allow the model to learn about the new subject (new vocabulary). During the training stage this new vocabulary is represented as follows: "a close-up photo of a S^*" or "a good photo of a S^*". The association between user-specified images (usually 3-5 images) and the new unique vocabulary (S^*) is learned by minimizing the reconstruction loss from the latent diffusion model (LDM loss) [36]. This optimization process is known as "Text Inversion". In contrast to DreamBooth, the primary goal of Text Inversion is to learn new text embeddings that correspond to the target concept, such that some aspects of the text are faithfully reflected in the synthesized image. DreamBooth focuses more on diverse image generation.

Although DreamBooth and Text Inversion can generate personalized images, their efficiency and utility are constrained by the need for numerous reference photos and complex training. Han *et al.* developed HiPer [66] for text-to-image personalization using text embedding decomposition and one target image. First, they investigated the CLIP embedding space for the prompt processing. Based on their findings, while the initial part of the CLIP embedding space (the first few dimensions) corresponds to low-level features such as colors and texture, the tail part (the last few dimensions) corresponds to objects and concepts. Therefore, by preserving the tail part of CLIP embeddings, identity information related to the target domain can be learned as well. For training, the last N tokens from the input prompt are selected (HiPer embedding) and concatenated with a personalized embedding as the condition for the pre-trained text-to-image generation.

B. Text-to-Motion Generation

Numerous studies have been undertaken for human motion generation. In this task, synthesized avatars are expected to maintain a fluent motion across the frames and be in sync with the input audio. Conventional approaches rely entirely on motion capture systems [67] and hardware [68] for the development of realistic human motion models. Later studies approached this task by employing music [69] and text [35], [62], [70] as input conditions to construct versatile and realistic human motion. The generated motions are expected to create aligned movements by being driven only by the input condition. The primary challenge in this task is connecting linguistic concepts to motion animations.

To achieve a cross-modal understanding between motion sequences and conditions, learning a rich joint space is crucial. Ahuja *et al.* developed Joint Language2Pose (JL2P) [71] by learning a joint embedding space for the text description and pose animation. This joint embedding space is created using a sentence encoder and pose encoder for the language and motion processing respectively. Since the created joint embedding space plays a crucial role in the quality of the generated motions, a joint translation loss is employed to make sure that the corresponding text and motions are close to each other. After the latent space construction, a pose decoder

generates a sequence of pose motions. The generated motion is highly dependent on how exact and detailed the input prompt is. To reduce the impact of this limitation, Transformer-based models have been further investigated. Petrovich *et al.* proposed the Action-Conditioned Transformer VAE (ACTOR) [70] for human motion generation. ACTOR uses a Transformer architecture, enabling the processing of long-range sequences respecting the expected relations between the body parts. A key component of the ACTOR architecture is the use of positional encoding in the decoder of the Transformer. This prevents the generated motions from regressing to the mean pose.

In the aforementioned methods, postures are generated based on a single action label such as "stand up" or a text description such as "a man walks a few steps". Being confined to a single action makes the output less realistic and lacks critical details. TEMOS [62] addresses this issue by extending the ACTOR network [70] using the pre-trained DistilBERT [72] model. Unlike the previous studies that depend on the motion sequences of the previous step to generate the next sequence of motion, TEMOS generates motion sequences in one shot, resulting in realistic motions and preventing static pose sequences.

Despite generating multiple human actions, TEMOS-synthesized avatars cannot reflect stylized motions (Fig. 8-(a)). To align the style of the synthesized moving avatar with the input condition, Tevet *et al.* developed MotionCLIP [63], a Transformer-based text to 3D motion generation method that is capable of elaborating textual data. This model injects the visual perception of CLIP into human motion generation. By using CLIP, MotionCLIP can generate actions not seen during training and exhibit abstract language capabilities. This endows the model with the capability to synthesize 3D avatars performing a set of relevant actions that are driven by the input prompt without that action explicitly being described. For example, by receiving the phrase "Swan Lake", it can generate a sequence performing ballet dance (Fig. 8-(b)).

Youwang *et al.* proposed CLIP-Actor [64] for text-to-motion generation. Similar to MotionCLIP [63], CLIP-Actor takes advantage of the joint text-vision space of CLIP for 3D human motion generation. However, this work put a special emphasis on aligning the style of the generated avatar to the given text prompt. For example, the synthesized moving avatar from the "walking Steve Jobs wearing blue jeans" is expected to not only perform walking but also resemble Steve Jobs while wearing blue jeans. In order to achieve this, two main components were further proposed: spatio-temporal view augmentation and mask-weighted embedding attention. Using these two components, CLIP-Actor can increase the similarity between the input text and the 3D motion avatar for realistic texture and shape generation.

Generating 3D motion that accurately reflects the action and style described in the given text prompt is time-consuming [65]. This further assumes more importance when the majority of the motion synthesis methods generate avatars in an autoregressive manner. To overcome this challenge Hong *et al.* proposed AvatarCLIP [65], by leveraging the strength of a large-scale pre-trained model, resulting in a zero-shot text to 3D motion generation model that is capable of building customized avatars in style and motion (Fig. 8-(d)). AvatarCLIP has three main steps in the process of stylized motion generation. First, the 3D human geometry is generated by the shape VAE network based on the input text description. Next, the 3D shape is improved by aligning the textures for the avatar using the volume rendering model. In the final stage, the CLIP-guided motion synthesis module generates 3D motion sequences based on the CLIP text encoder.

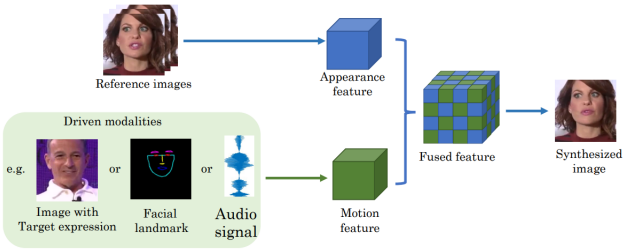


Fig. 9: An overall architecture for talking face generation [74].

Although promising results have been achieved by previous studies in text-to-motion generation, the existing methods often fail to handle complicated motion descriptions. For example, given the "shaking head and waving hand" description, the generative model is expected to have control over different body parts to accurately synthesize the avatar performing head shaking and hand waving simultaneously. Inspired by the progress seen by diffusion models [32], especially in text-to-image generation, Zhang *et al.* proposed MotionDiffuse [35] for text-to-motion generation. In MotionDiffuse, motion sequences are generated through a series of denoising steps. Since a different amount of noise is added to the sequences through diffusion, the model gradually learns complex distributions and variations in human actions. By leveraging a diffusion model and large language model, realistic and diverse motion animations are generated by MotionDiffuse (Fig. 8-(e)). More recently, MoFusion [73] surpassed the previous state-of-the-art by generating longer arbitrary-length motion animations, conditioned on both text and audio. MoFusion introduces a novel time-varying weight schedule to DDPM for temporally and semantically aligned output. More specifically, this network is conditioned on Mel Spectrogram (audio) and CLIP Tokens (text) to control the generated motion animations.

IV. Audio to Visual Content Generation

Audio-to-video generation aims to synthesize high-resolution and photorealistic videos conditioned by audio. To better discuss the methods in this area, we divide this section into talking face and dance choreography generation.

A. Talking Face Generation

It is hard to obtain talking faces that clearly express specific speech information since the dynamic deformation of the facial region is subject-specific and speech-dependent [75]. As a result, audio features, facial landmarks, and sample identity frames are usually used jointly for talking face generation (Fig. 9). The early research focused on generating video by

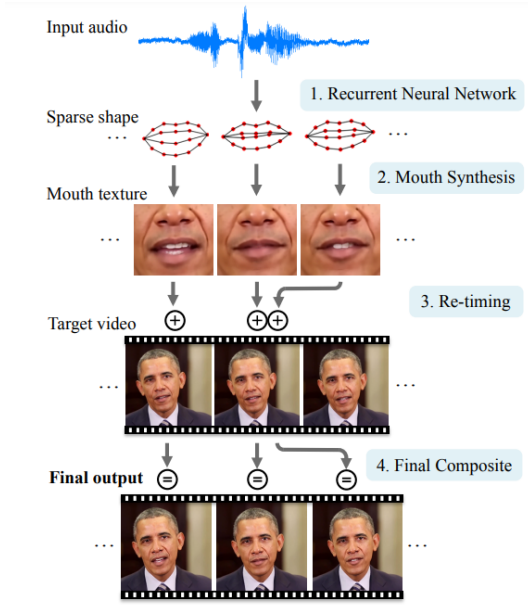


Fig. 10: The Synthesizing Obama [77] architecture.

mapping audio to mouth key points. ObamaNet [76] is among the first neural network-based models developed for audio-to-talking face generation using mouth key points. Using a Text-to-Speech network, text descriptions are first transformed into audio representation. Next, with the Time-delayed LSTM, mouth landmarks are extracted in a way that aligns with the given audio sequence. Last, a sequence of video frames is generated using a pix2pix network conditioned on mouth key points. Suwajanakorn *et al.* proposed the "synthesized Obama" [77] as a talking head system using a relatively similar approach to ObamaNet [76]. This method utilizes a Recurrent Neural Network (RNN) trained on Obama's videos for mapping audio sequences to the corresponding mouth shapes. By leveraging this learned mapping, the system can generate realistic lip sync for arbitrary audio inputs. Although these studies have achieved promising results, their generalization ability is confined to a single identity (Fig. 10).

Many existing methods in talking face generation rely on mouth key points, which require accurate phoneme labels within millisecond time steps as input. They fail to accurately morph the lip movements of an arbitrary person in a dynamic and uncontrolled environment, leading to major chunks of the video being out of sync with the audio description [76], [77]. To address this limitation and reduce out-of-sync lip motions, Si *et al.* proposed speech2video [78] to effectively extract visual attributes from the input audio. For this purpose, a cross-modal distillation network comprised of a student-teacher strategy has been developed. These extracted intermediate features are further utilized to train a GAN for talking face generation.

To improve coherency and enforce temporal stability in the synthesized videos, depth maps have been used in some studies. Constructing 3D features removes dependence on identity information from the source image and enhances generalization across different identities [79]. Although 3D depth information improves the synthesized results, 3D geometry annotations are often not available for video generation tasks and annotating them can be excessively expensive. To

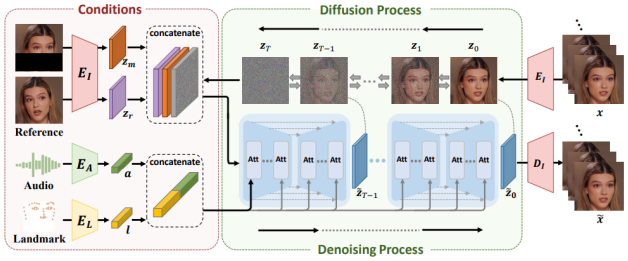


Fig. 11: The architecture of DiffTalk.

reduce the impact of this problem, while taking advantage of the depth map in face generation, Hong *et al.* proposed an automated method for dense 3D geometry (depth) extraction from face videos [79] in a self-supervised manner. Leveraging the extracted depth maps and depth-aware GAN (DaGAN), identity and pose-preserved talking face videos are generated.

The subject-related and speech-related information are intertwined in audio, making it challenging to develop speaker-independent video generation. Most recent research has focused on addressing this challenge by developing mechanisms to extract identity-related information from audio. Prajwal *et al.* proposed an automatic GAN-based lip sync expert, known as Wav2Lip [6], for lip movement prediction from speech. This framework can be considered the first speaker-independent model in talking face generation. To eliminate out-of-sync lip movements, a trained discriminator is utilized and fine-tuned on the training data. The generator consists of three main components that are necessary to extract audio information and maintain identity: the identity encoder, speech encoder, and face decoder. Wav2Lip can synthesize videos aligned and lip-synced to any arbitrary audio. In a similar manner, Zhou *et al.* introduced MakeItTalk [80], a speaker-aware talking head video generation framework using audio segments and reference images as input. MakeItTalk has two main steps: a lip-sync module and a face animation module. Lip-sync is trained to map audio sequences to their corresponding visual speech representations. The representations are then used to refine the generated lip sequences and produce a talking-head video using Image2Image translation. MakeItTalk reconstructs videos containing facial expressions and eye blinks.

Over the past years, researchers have overcome different challenges in talking face generation. However, the main challenge of generating a talking face video with natural expressions and no additional guidance remained unresolved. Recent improvements in diffusion models [32] yield promising results for end-to-end talking head generation [81]. DiffTalk [33] and Diffused Heads [81] are among the first such models.

DiffTalk [33] is a diffusion-based method for talking head generation that operates on the latent space. This model has three inputs (reference, ground truth and masked frames) for training and is conditioned on facial landmarks and audio representation. To enhance the model generalization ability, the mouth region landmarks are removed as shown in Fig. 11. Based on the reported results, using face landmarks and an identity reference frame is crucial for having a personality-aware generative model. With the assistance of diffusion models, DiffusedHeads [81] can synthesize talking sequences using only an identity image and an audio sequence. During

Fig. 12: A comparison of some representative talking face generation methods.

training, each frame is generated in a one-at-a-time manner. DiffusedHeads is conditioned on motion frames, an identity frame, and audio representation produced using a pre-trained audio encoder [82]. This model can generate natural talking faces (with head movements and eye blinks) while maintaining the background of the identity frame during the generation process.

In Fig. 12, we compare some renowned audio-to-talking face generation models using a subject chosen from the HDTF dataset [83]. Wav2lip, while offering a functional solution, it shows noticeable drawbacks. The generated frames often exhibit blurriness, and lip movements are frequently misaligned. MakeItTalk, on the other hand, presents a noticeable improvement in image quality and lip-audio alignment, compared to Wav2lip. However, similar to Wav2lip, MakeItTalk struggles with rendering the inside of the mouth and detailing the teeth during speech animation. DiffTalk, leveraging a stable diffusion model, surpasses the previous methods by producing high image quality and achieving substantially improved lip-audio alignment. The utilization of a stable diffusion model and additional conditions (face landmarks) has contributed to the reduction of jitter and enhanced the overall visual fidelity. However, there is still room for further refinement as some generated frames are not fully aligned with the corresponding audio and ground truth frames.

B. Dance Choreography Generation

Dance choreography generation involves synthesizing motion sequences that follow the beats and rhythms of the music while maintaining temporal consistency across motion frames. In earlier studies, traditional methods such as clustering [84], motion correlation coefficients [85], and statistical models [86] were employed for music-to-dance generation. The motion sequences generated by conventional methods lack realistic choreographies and do not emulate the complex human dancing abilities. In recent years, deep learning has emerged as a promising alternative to conventional approaches and is now

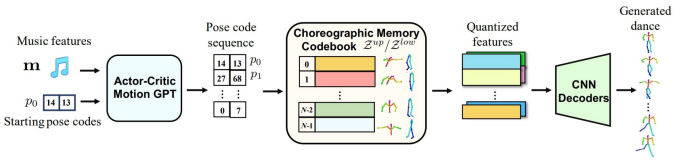


Fig. 13: The architecture of Bailando [90].

extensively used for dance movement generation. Alemi *et al.* developed GrooveNet [87], a deep neural network model based on factored conditional restricted Boltzmann machine and RNN, which is capable of generating dance movement sequences driven by audio.

Although deep learning methods can synthesize moving avatars, their movements are often irregular. To overcome this shortcoming and improve the pose dynamics estimation, the attention mechanism has been widely used by the existing generative methods. For instance, Kao *et al.* developed a generative-attention-based network for synthesizing the movement of the skeleton of a violinist playing a particular piece of music [88]. This method first extracts the Mel-Frequency Cepstral Coefficients (MFCC) features of the audio. Then, the sequence of virtual skeletons playing the violin are constructed using an encoder-decoder network. To further improve the accuracy of the predicted posture, Ren *et al.* developed a pose generator integrated with the GRU network [89]. Using a graph-based representation of the body joints, the Spatial-Temporal Graph Convolutional Network (ST-GCN) is employed as a pose generator. ST-GCN learns spatial and temporal relations of a given sequence and produces a better alignment with the movements of the skeleton. Similar to previous methods, audio features are computed using an audio encoder. These features, extracted with the assistance of ST-GCN, are transformed into dancing skeletons. To generate realistic dancing avatars from skeletons, the pix2pixGAN method has been widely used.

Most prior works use generative models with many networks and sub-networks, making the overall architecture complex [91]. Despite the multiplicity of the conditioning layers employed in their architectures, the synthesized dance movements are often repetitive and misaligned with the music rhythm. To generate diverse and aligned dance movements while maintaining the simplicity of the overall architecture, Transformers have been used for the music-driven dance generation. With the self-attention mechanism, Transformers can capture longer sequences better than conventional sequence-based models such as RNN and LSTM. Li *et al.* developed a Two-Stream Motion Transformer (TSMT), consisting of audio-stream and pose-stream Transformers, to extract pose and audio information [92]. TSMT uses a fusion module to combine the extracted features for motion prediction.

Robust and accurate pose estimation for dancing skeletons plays a pivotal role in bridging the gap between the music and visual modalities. Nevertheless, due to the stochastic nature of dance, developing a spatiotemporal balanced dance generation model is challenging. Siyao *et al.* introduced Bailando [90] for spatially and temporally coherent dance movement generation using a choreographic memory and an actor-critic Generative Pre-trained Transformer (GPT). Utilizing the music and a

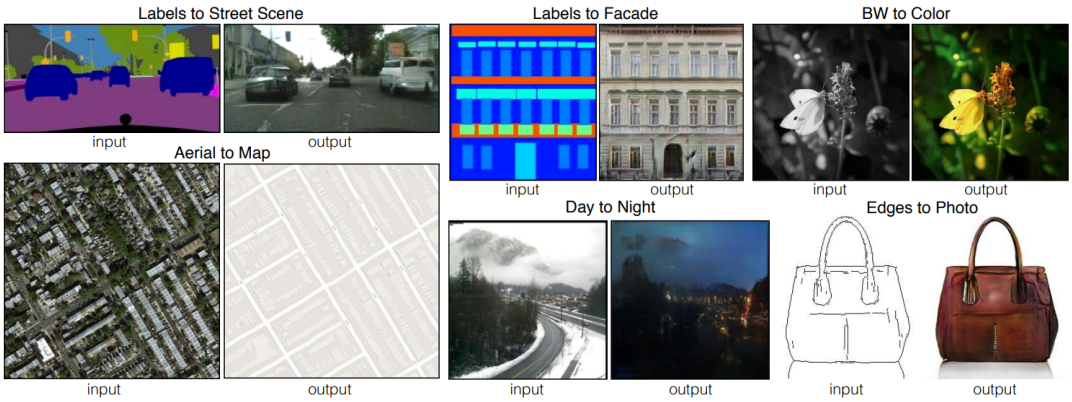


Fig. 14: Some examples of the cross-domain image-to-image translation task [94].

starting pose code, actor-critic GPT produces a sequence of future pose codes, each representing a dancing pose. It is important to note that the dance pose codes are generated for the upper and lower bodies separately (Fig. 13). This can improve the variety of synthesized motions in dance generation during inference. These vectors are embedded into a quantized latent space using VQ-VAE [59]. By applying a CNN decoder to the retrieved quantized features, the final 3D dance sequences are constructed.

Although the research conducted in this area has eventually managed to generate dancing skeletons with poses that match the corresponding music, the controllability of gestures has not received much attention. More recently, to mitigate this issue, Alexanderson *et al.* proposed a diffusion-based model for style controlled motion generation based on the input audio and an optional style reference [91]. The advocated architecture is based on Diffwave [93], an audio synthesizer model. This model has been trained on audio-to-motion and audio-to-dance generation datasets; hence, it is applicable to both tasks. To enhance the controllability of the synthesized dance, Tseng *et al.* proposed Editable Dance Generation (EDGE), which is a Transformer-based diffusion model. This model not only modifies and controls the dance poses based on a user's preference but also generates dance sequences of arbitrary length with fewer jitters. EDGE makes use of a contact consistency loss to capture the physical ground contact behavior so as to control the movements of the feet of the virtual dancing skeletons. This results in even more realistic motions.

V. Visual Content Generation Based on Other Modalities

Visual content generation is not restricted to text and audio conditions. We can guide the process by using other modalities, such as semantic maps and sketches for this task. In this section, we investigate the existing cross-domain image-to-image translation and sketch-based vision generation methods, summarising the underlying principles, challenges, and applications.

A. Cross-domain Image-to-Image Translation

The task of cross-domain image-to-image translation is to transfer images from a source domain to a target domain, while preserving the content of the source image [95]. Some examples are shown in Fig. 14. Based on the data used for training, the existing methods can be categorized into supervised and unsupervised approaches [96].

Common supervised approaches include Pix2Pix [94], Pix2PixHD [97] and SPADE [98]. In these approaches, every image in the source domain has a corresponding image in the target domain. Pix2Pix is a GAN-based network comprising U-Net as the generator and patchGAN as the discriminator. This network generates new images by processing the provided semantic maps. Although semantic maps carry considerable information about the structure of the target image, the images synthesised by Pix2Pix are not always of high quality and resolution [94]. Later, Pix2PixHD [97] developed multi-scale generators and discriminators to address this issue. The images generated by Pix2PixHD are of resolution 2048 × 1024. SPADE uses an adaptive normalization layer conditioned spatially for image synthesis. Given an input semantic layout, this new normalization mechanism captures the semantic information more effectively than previous normalization layers such as InstanceNorm, resulting in high-resolution, high-fidelity image generation. The normalization is performed in a spatially-adaptive manner based on the semantic content of each section of the image, which is very much different from utilizing the same normalization parameters throughout the entire image [98].

On the other hand, unsupervised approaches frequently base their modeling procedure on a shared latent space. Using unpaired training data, models in this task attempt to find an accurate mapping between a source domain and a target domain [96]. Kim *et al.* introduced DiscoGAN, which has two generators and two discriminators. DiscoGAN learns to map from one domain to another and vice versa without using labeled data [99]. Unsupervised approaches facilitate the acquisition of cross-domain relations and transfer characteristics such as style. Chen *et al.* proposed Vector Quantized Image-to-Image Translation (VQ-I2I) [100] for unconditional image-to-image translation. VQ-I2I utilizes the vector quantization approach and consists of three main components: content encoder, domain-specific style encoders, and domain-specific decoders. First, the VQ-based content encoder transforms the input semantic map into a vector-quantized codebook. Next, style and domain-specific features are learned using the discrete features extracted by the domain-specific encoder.

This allows for inter-domain and intra-domain translation. For example, it can be used to change the color of the eyes or turn an image scene from winter to summer.

Recently, diffusion models have surpassed the state-of-the-art in common evaluation metrics by generating higher quality and more realistic visual content. Saharia *et al.* proposed Palette, an image-to-image translation diffusion-based model, by employing the standard diffusion technique with the U-Net backbone. This method demonstrates that a wide range of image-to-image translation tasks such as colorization, inpainting, and JPEG restoration could be addressed by diffusion models [101]. The diffusion models learn to synthesize content by following the Markov chain of a denoising process. To learn a richer model of the source content for transfer to the target domain, Sasaki *et al.* proposed UNpaired Image Translation with DDPM (UNIT-DDPM) [102] by applying a dual-domain Markov chain for the diffusion process. This mechanism approximates the data distribution of the source and target by denoising in a joint space.

Considering the efficiency of using a shared latent space for training an image-to-image translation diffusion model, Li *et al.* proposed a new approach using the Brownian Bridge Diffusion Model (BBDM) [103] for image-to-image translation. Instead of the standard conditional generation process, this method interprets translation as a stochastic Brownian bridge process and directly learns to bridge two domains using a bidirectional diffusion. Similar to the latent diffusion model [36], the stochastic Brownian bridge process is deployed in the latent space of VQGAN. In contrast to standard diffusion models, which start from an image and continue with Gaussian noise at various scales, the stochastic Brownian bridge process begins from the source image and sets the target source as the destination. In this way, the denoising process gradually learns to directly map across domains.

B. Sketch-based Image Generation

Among the existing cross-domain image-to-image translation tasks, sketch-to-image generation has received great attention due to its wide practical applications. This task seeks to synthesize a detailed image corresponding to a particular hand sketch. Earlier methods such as Sketch2photo [104] and Photosketcher [105] heavily rely on the quality of extracted features and post-processing techniques such as graph cut compositing [106]. With the emergence of deep learning and generative models, this area has experienced a rapid uplift. Before continuing with the explanation of this section, it is important to note that the performance and generalization capability of deep learning-based models are largely dependent on the quality and quantity of the training samples. Due to the limited number of samples for sketch-to-image generation and the labour-intensive nature of obtaining the associated drawing for each image, the edge map is frequently used for sketch-to-image generation.

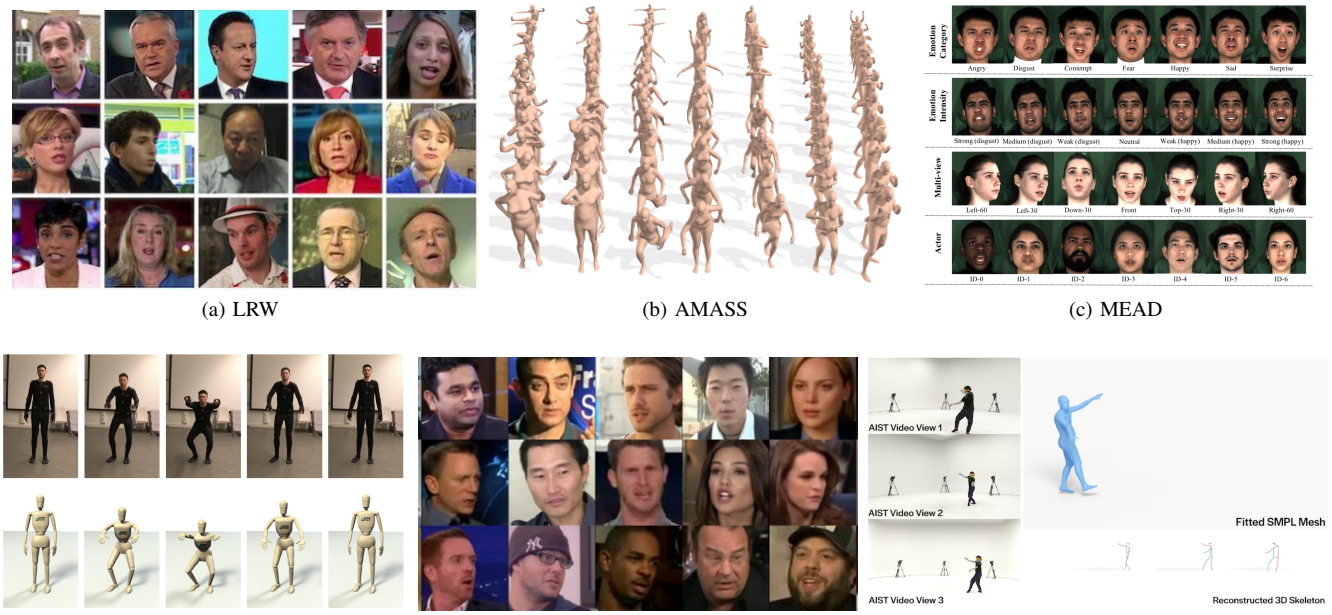
Chen *et al.* proposed SketchyGAN [106] for image generation based on sketches. Due to a limited number of samples in training, edge maps are constructed via holistically nested edge detection. Next, to make edge maps sketch-like, data augmentation is introduced. Based on the reported results, increasing the number of training data samples results in improved generation performance. Further, another contribution of this work is the introduction of a Masked Residual Unit (MRU) to both the generator and discriminator. This block contributes to the information flow of the network by processing the feature maps of image edges to condition the generated images.

Similarly, Li *et al.* proposed a Conditional Self-Attention GAN (CSAGAN) [107] for sketch-to-image generation. They concentrated specifically on face sketch-to-image generation. Generating face images from sketches is challenging, especially if the conditional line segments are incomplete. To encourage the model to capture the face structure and to learn the relations between distinct regions, the self-attention technique is employed in the generator. The generator is an encoder-decoder architecture, which takes advantage of MRUs to share information across the network. This mechanism, integrated with self-attention, can assist the model in learning the local and global context of faces.

Due to the lack of paired image and sketch datasets, Bhunia *et al.* proposed the use of unlabeled images to boost the efficacy of sketch-to-image generation [108]. This study developed a semi-supervised framework, consisting of two major components that jointly train photo-to-sketch generation and Fine-Grained Sketch-Based Image Retrieval (FG-SBIR) models. Photo-to-sketch generation can be considered the core of this framework, as it generates sketch pairs for unlabeled images and facilitates sketch-to-image generation. Since there is a risk of low quality and misaligned sketch generation, a discriminator-guided mechanism is leveraged to guide the sketch generator in producing high-quality sketches.

As mentioned earlier due to the scarcity of sketch-image pairs, edge maps are utilized extensively for this task. Edge maps are primarily boundary representations of an image, but human drawings vary in style and precision. Given the same description such as “a cat sitting on a bench”, each person creates a unique sketch that is semantically comparable. It is essential to train sketch-to-image generation models using sketches rather than edge maps in order to preserve the diversity and accuracy of the generated results. However, using only sketch-image pairs has its own set of obstacles. For instance, due to limited diversity in sketch-image pairs, SketchyGAN [106] can only generate up to 50 image categories.

To overcome these limitations, Koley *et al.* proposed “Picture that Sketch” [109] for photorealistic sketch-to-image generation. This model is an encoder-decoder model comprised of an autoregressive sketch mapper as the encoder and StyleGAN as the decoder. The proposed autoregressive sketch mapper has been trained to map a given sketch to its corresponding image StyleGAN latent space. Next, using a fine-tuned StyleGAN model and the generated vector from the sketch mapper, a photorealistic image of a given sketch is generated. To control the training and ensure alignment between the ground truth and the synthesized image, the reconstruction loss and fine-grained discriminative loss are employed.



(a) LRW (b) AMASS (c) MEAD

(d) KIT Motion Language (e) Voxceleb (f) AIST++

Fig. 15: Examples of the LRW [24], AMASS [115], MEAD [27], KIT Motion Language [119], Voxceleb [26] and AIST++ [69] datasets.

TABLE I: Datasets for cross-modal content generation.

Task	Datasets	Detail
Speech-to-Video	LRW [24]	450,000 video clips
	LRS [25]	118,116 video clips
	Voxceleb [26]	22,496 video clips
	GRID [110]	34,000 video clips
	CREMA-D [111]	7,442 video clips
	MEAD [27]	281,400 video clips
	MOSEI [112]	23,453 video clips
	CMU-MOSI [113]	2,199 video clips
	LLP [114]	11,849 video clips
	HDTF [83]	300 video clips
Text/label-to-Motion	AMASS [115]	11,451 motion sequences
	HumanAct12 [116]	1,191 motion clips
	BABEL [117]	40 hours of mocap data
	HumanML3D [118]	14,616 motion sequences
	KIT Motion Language [119]	3,911 motion clips
Speech-to-Motion	Talking with hands [120]	16.2M video frames
	Trinity Speech Gesture(TSG) [121]	4 hours motion videos
Music-to-Dance	GrooveNet [87]	0.38 hours dance videos
	Dance w/Melody [122]	1.6 hours dance videos
	AIST++ [69]	5.2 hours dance videos
	MMD [123]	19.91 hours dance videos
	FineDance [124]	14.6 hours dance videos
Text-to-Image/Video	CUB [125]	11,788 images
	MS-COCO [126]	328k images
	ImageNet [127]	14k images
	LAION [128]	400M image-text pairs
	UCF-101 [129]	13,320 video-label pairs
	CelebA-Dialog [130]	202,599 image-text pairs
	Conceptual 12M [131]	12M image-text pairs

VI. Datasets And Evaluation Metrics

This section introduces the datasets and evaluation metrics used for the visual content generation. It helps in understanding the strengths and limitations of the current datasets and metrics, identifying potential gaps that need to be addressed.

A. Datasets

Given the crucial role of data in the development of a robust generative model, we summarize the commonly employed datasets for cross-modal content generation in Table I. In particular, we have provided datasets employed in both text-to-visual and audio-to-visual content generation. Further, we would like to note that some research works such as DALL-E [43] have collected their own internal dataset in addition to utilizing publicly available datasets. We have to exclude this dataset from this table, as we do not have access to them.

Most previous literature addressed datasets related to image-to-vision and text-to-vision synthesis, while a little amount of work focused on audio (speech and music) to visual content (image and video). To fill the gap, we have put a special focus on audio-to-visual content generation datasets. For simplicity, we have categorized data into four categories: speech-to-video, text/label-to-motion, speech-to-motion, and music-to-dance generation, respectively. Some examples of the datasets are shown in Fig. 15.

To assist researchers in choosing a suitable dataset or determining gaps and improving them, we briefly compare commonly utilized datasets amongst their counterparts for each category. We would also like to note that choosing between datasets largely depends on the specific research question or application. In some cases, combining insights from multiple datasets could provide a more robust and comprehensive understanding of patterns and relations among data for training a model.

One of the most well-known datasets for the talking face generation is LRW [24], which has served as a benchmark for

TABLE II: Typical evaluation metrics used for cross-modal visual content generation.

Task	Metric	Description
Video and Image Synthesis	Inception Score (IS) [132]	Evaluating the visual quality
	Frechet Inception Distance (FID) [133]	Evaluating the visual quality
	Frechet Video Distance (FVD) [134]	Measures the fidelity and diversity of generated samples
	Structural Similarity Index Measure (SSIM) [135]	Measures the similarity between two images
	Peak-Signal-to-Noise Ratio (PSNR)	Measures the visual quality between two images
Dance Choreography Generation	Beat Alignment Score (BAS) [73]	Measures similarity between motion and audio beats
Motion Generation	Motion-retrieval precision (R-precision) [136]	Calculates the text and motion top 1,2,3 matching accuracy
	Reconstruction Accuracy [137]	Average joints positions distance and root trajectory distance
Text-to-Video	CLIP Similarity (CLIPSIM) [138]	Average CLIP similarity between video frames and text

a variety of related tasks. This dataset was collected by BBC, ensuring a professional production quality. Voxceleb [26] contains a relatively higher number of videos but they are collected online. Videos can vary in quality, but most of them are of acceptable quality. HDTF [83] is a relatively new dataset specifically captured for audio-to-talking face generation. It has focused on high-quality data collection. Its videos are longer, compared with LRW and Voxceleb. CREMA-D [111], MEAD [27] and GRID [110] are captured under a controlled environment, presenting well-structured, consistent and high-quality videos. These attributes make them appropriate for talking face generation, lip reading, emotion, and expression analysis.

AMASS [115] stands out for its extensive collection of data from various sources and its utility in 3D body reconstruction. Humanact12 [116] and BABEL [117] also contain acceptable motion sequences. However, if the intersection of language and motion is the primary interest, then the BABEL dataset can be a better option.

For tasks that need precise hand gesture predictions based on a spoken language, "Talking with Hands" [120] would be the preferred choice. However, if the objective is to understand or generate broader body motions in response to speech, then TSG [121], with its comprehensive capture of full-body gestures, becomes more relevant. In the context of music-to-dance generation, FineDance [124] is a relatively newer dataset and addressed prior issues in terms of the sample size, diversity in dance movements and annotation accuracy. Information related to skeletons is reported in 3D space, providing more control and realism to generated dancing avatars.

MS-COCO [126] is a widely employed dataset for image generation tasks. It has $328k$ images across 80 categories, offering a broad foundation for diverse applications. Although employed for image generation, it was primarily created for object detection and segmentation. On the other hand, LAION [128] and Conceptual12M [131] present a rather specialized dataset, merging a vast number of images with their corresponding textual descriptions. With their extensive volume, they promise a robust model training opportunity. CelebA-Dialo [130] offers task-specific text-to-image generation data. It is suitable for generating human faces from detailed textual prompts describing expected facial attributes.

We further outline several limitations associated with the current datasets to assist researchers in identifying and understanding the challenges in cross-modal visual content generation.

- Although the quality of current collected video and audio has improved as a result of advanced equipment, they still suffer from being of limited size and diversity, especially among music-to-dance generation datasets. This restricts the ability of a trained model in generalization and robustness.
- Cross-modal datasets often suffer from flawed annotations which limits the ability of a model in capturing information about the relationships between different modalities and accurate mapping from one modality to another. This affects the alignment between music to dance movements and the synchronization between audio and lip movements of 3D avatars.
- The last and most important issue associated with cross-modal datasets that we want to point out is privacy and ethical concerns. This issue makes data gathering in this field challenging as ensuring the privacy and consent of individuals involved in data gathering is important.

B. Evaluation Metrics

Different assessment criteria for generative models have been established in the field of visual content synthesis. Table II briefly describes these metrics. In addition, we summarize the evaluation results of some popular methods in cross-modal visual content generation in Table III. Although we have specified categories such as dance choreography generation and text-to-video, it is noteworthy to mention, that metrics provided in the video and image synthesis group can be utilized for the majority of the tasks involved in generating visual content.

TABLE III: A comparison of representative methods in cross-modal visual content generation.

Task, Dataset	Model	Evaluation Metric
		FID↓
Text-to-Image (MS-COCO [126])	Cogview [29]	27.10
	DALLE [43]	17.89
	GLIDE [50]	12.24
	DALL-E2 [57]	10.39
	Stable Diffusion [36]	12.63
		R-Precision↑
Text-to-Motion (HumanML3D [118])	MotionCLIP [63]	0.0029
	AvatarCLIP [65]	0.0002
	MotionDiffuse [116]	0.491
		SSIM↑
Audio-to-Talking face (HDTF [83])	MakeItTalk [80]	0.544
	Wav2Lip [6]	0.761
	DiffTalk [33]	0.950
		FID↓
Sketch-to-Image (Sketchy dataset [140])	Pix2Pix [94]	33.4
	PhotoSketch [105]	25.7
	FG-SBIR [108]	8.9

This deficiency led some works [43], [52] to apply qualitative analysis such as human evaluation to assess the quality of the generated visual content. The primary drawbacks of this qualitative metric are irreproducibility and the extremely subjective nature of them [139]. This is considered a major limitation in cross-modal learning.

As mentioned earlier, metrics such as Frechet Inception Distance and Peak-Signal-to-Noise Ratio are often employed across various cross-modal generation tasks. Although these metrics have been employed frequently and can be considered a good starting point for comparing outcomes obtained from novel methods to previous state-of-the-art, they are not completely reliable. Some research [81], [139] reported how PSNR penalizes their results and that there is a gap between subjective scores in terms of human perception and the objective measures in this metric.

Each modality has unique characteristics and using the same evaluation metrics may not fully capture the quality, diversity, or semantic coherence of the generated content. For instance, in audio-to-talking face generation, evaluation metrics require focusing on multiple aspects including but not limited to the quality of the generated video, the incorporation of realistic eye blinking during speaking, identity preservation, and audio-lip synchronization [74].

This is particularly important for recent tasks such as text-to-video generation. Generated videos should not only accurately reflect the input prompt but also contain spatiotemporal consistency across synthesized frames. This task requires further investigation as the majority of research on this topic [53], [54] either utilized human evaluation or extended text-to-image metrics for assessing the generated visual content. As a result, developing a task-specific and interpretable evaluation metric is highly encouraged.

VII. Challenges And Future Directions

Diffusion models are predominantly employed in various disciplines. They have demonstrated a significant improvement over the state-of-the-art methods. These models produce high-fidelity visual content, while avoiding the network collapse problem. Diffusion models are likelihood-based models and go through an iterative process to generate samples. This might lead to excessive use of computing resources [36]. Several studies proposed executing denoising steps in the latent space of images, while ensuring realistic and diverse visual content synthesis [36], [57], [58]. Developing models using a latent representation for the sampling process is highly encouraged to balance image quality and computational costs.

Despite the visual fidelity of the output synthesised by diffusion models, their capacity is limited to the task they have been trained on. It is efficient to develop a multi-task multimodal network, that can process data in different modalities for generating images, 3D samples, text and etc. Versatile Diffusion (VD) [141] is a sample of this concept. Although VD is capable of processing text-to-image and image-to-text, further research on this topic is required to tackle a wider range of modalities.

Despite demonstrating improvements in the visual content generated by diffusion models, a further study of other generative models is encouraged. Several recent studies have shown realistic and aligned synthesized visual content by utilizing Transformers [142] and GANs [79], [143] as the primary generative model. In comparison to diffusion models, GANs require a single forward step, which makes them faster in training.

In spite of the efficiency in training speeds and strong alignment between the input and the synthesized results, these models (especially GANs) are behind the current state-of-the-art in terms of generation performance. This should not prevent but rather encourage further research on these generative models for cross-modal visual content generation. As Wang *et al.* demonstrated, integrating GAN with a diffusion model, *e.g.,* Diffusion-GAN [144], can address generative adversarial network training limitations (such as slow convergence and mode collapse [39],) while further improving the diversity and quality of generated images. This undoubtedly deserves further investigation to determine its applicability to cross-modal tasks.

Transformers, on the other hand, have demonstrated exceptional performance in natural language processing tasks and as condition encoders (*e.g.,* text and audio encoders) in multi-modal generation tasks. However, there is still room for further investigation of this architecture in cross-modal visual content generation. It is encouraged to explore the performance of transformers as the main learning module in this domain. Their attention mechanism should be repurposed to best link the source domain to the target domain.

Furthermore, in order to synthesize realistic coherent motion in videos, generating arbitrarily long sequences requires more attention. Some work addressed this using a Transformer-based decoder [62], or by conditioning on a sequence of distinct text prompts at each time step [55]. However, the majority of the video synthesis methods in the literature are restricted to generating fixed-length videos. This affects their ability to handle complex scenes and produce a realistic video frame generation. Further, if the generated visual content

includes a moving avatar or talking face, the ability to produce a dynamic pose and a relevant emotional expression in concert with the input modality (speech or text) is crucial. Currently, most of the generative models exhibit limited head movements, greatly impacting the realism of the synthesized content. This can be investigated further by enhancing joint latent space and cross-attention mechanisms.

VIII. Conclusion

Cross-modal content generation facilitates the synthesis of information from one modality to another, hence enabling the generation of expressive and controllable content. In this survey, we comprehensively reviewed recent visual content generation methods across different modalities. In particular, we grouped the research works based on their input modality into text-to-vision, audio-to-vision, and other modality-guided visual content generation. It is important to jointly investigate these modalities for visual content generation as they share common underlying methods. Understanding one domain well can facilitate the development of another domain.

In addition to reviewing the recently published methodologies, we presented a concise comparison of the current datasets and metrics. This information provides an insight into the limitations of the conventional assessment of visual content generation methods. The lack of reliable and task-specific evaluation metrics is one of the main shortcomings. Through the examination of recent studies, we have catalogued improvements and identified open challenges in this domain, such as the task of extending cross-modal visual content generation to multi-modality visual content generation.

The survey of cross-modal visual content generation summarised the current state of the art. By discussing the methods, datasets, and evaluation metrics, we have identified both the progress and remaining challenges. We hope this survey will help to promote the future research in the cross-modal visual generation domain and lay the ground for advancing the state-of-the-art in this field.

Acknowledgments

This work was supported in part by the EPSRC grant MVSE (EP/V002856/1).

References

[1] J. Agnese, J. Herrera *et al.*, "A survey and taxonomy of adversarial neural networks for text-to-image synthesis," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, 2019.

[2] C. Zhang, C. Zhang *et al.*, "Text-to-image diffusion model in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[3] R. Zhen *et al.*, "Human-Computer Interaction System: A Survey of Talking-Head Generation," *Electronics*, vol. 12, no. 1, 2023.

[4] T. Sha, Zhang *et al.*, "Deep Person Generation: A Survey from the Perspective of Face, Pose, and Cloth Synthesis," *ACM Comput. Surv.*, vol. 55, no. 12, 2023.

[5] C. Sheng, G. Kuang *et al.*, "Deep learning for visual speech analysis: A survey," *ArXiv*, vol. abs/2205.10839, 2022.

[6] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild," in *ACM International Conference on Multimedia*, 2020.

[7] J. Tseng, R. Castellon, and C. K. Liu, "EDGE: Editable Dance Generation From Music," in *CVPR*, 2023.

[8] R. Chellappa and R. Kashyap, "Texture synthesis using 2-d noncausal autoregressive models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1985.

[9] G. R. Cross and A. K. Jain, "Markov Random Field Texture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1983.

[10] L.-Y. Wei and M. Levoy, "Fast Texture Synthesis Using Tree-Structured Vector Quantization," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000.

[11] W. Xu and E.-J. Lee, "Face recognition using wavelets transform and 2D PCA by SVM classifier," *International Journal of Multimedia and Ubiquitous Engineering*, pp. 281–290, 2014.

[12] W. Guo, X. You *et al.*, "Dynamic Texture Synthesis via Image Reconstruction," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 1907–1911.

[13] S. M. A. Eslami, N. Heess *et al.*, "The Shape Boltzmann Machine: a Strong Model of Object Shape," in *IJCV*, 2013.

[14] K. Zhang, W. Wang, Z. Lv *et al.*, "Computer vision detection of foreign objects in coal processing using attention cnn," *Engineering Applications of Artificial Intelligence*, vol. 102, p. 104242, 2021.

[15] S. R. Indurthi, D. Raghu, M. M. Khapra, and S. Joshi, "Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model," in *15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 376–385.

[16] V.-K. Tran and L.-M. Nguyen, "Natural Language Generation for Spoken Dialogue System using RNN Encoder-Decoder Networks," in *Conference on Computational Natural Language Learning*, 2017.

[17] Z. Zhang, Z. Li, K. Wei *et al.*, "A survey on multimodal-guided visual content synthesis," *Neurocomputing*, vol. 497, pp. 110–128, 2022.

[18] E. Mansimov, E. Parisotto *et al.*, "Generating images from captions with attention," *CoRR*, vol. abs/1511.02793, 2015.

[19] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in *BMVC*, 2017.

[20] I. Goodfellow, Pouget-Abadie *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[21] W. Fang, F. Zhang, V. S. Sheng, and Y. Ding, "A method for improving CNN-based image recognition using DCGAN," *Computers, Materials and Continua*, vol. 57, no. 1, pp. 167–178, 2018.

[22] H. Zhang, T. Xu *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017, pp. 5907–5915.

[23] T. Xu, P. Zhang *et al.*, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018, pp. 1316–1324.

[24] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham: Springer International Publishing, pp. 87–103.

[25] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR*. IEEE, jul 2017.

[26] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*. ISCA, 2018.

[27] K. Wang, Q. Wu, L. Song *et al.*, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, pp. 700–717.

[28] A. Vaswani, N. Shazeer *et al.*, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[29] M. Ding, Z. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *NeurIPS*, vol. 34, pp. 19 822–19 835, 2021.

[30] S. Naveen, M. S. S. Ram Kiran *et al.*, "Transformer models for enhancing AttnGAN based text to image generation," *Image and Vision Computing*, vol. 115, p. 104284, 2021.

[31] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Interspeech*, 2019, pp. 3465–3469.

[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[33] S. Shen, W. Zhao *et al.*, "DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation," 2023.

[34] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," in *CVPR*, 2023.

[35] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *arXiv preprint arXiv:2208.15001*, 2022.

[36] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[37] Y. Cao, S. Li, Y. Liu *et al.*, "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT," *ArXiv*, 2023.

[38] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the Generative Learning Trilemma with Denoising Diffusion GANs," in *ICLR*, 2022.

[39] A. Sauer, T. Karras *et al.*, "StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis," *ICLR*, 2023.

[40] H. Zhang, T. Xu *et al.*, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE TPAMI*, vol. 41, no. 8, pp. 1947–1962, 2018.

[41] A. Raganato *et al.*, "An analysis of encoder representations in transformer-based machine translation," in *EMNLP Workshop on BlackboxNLP*. The Association for Computational Linguistics, 2018.

[42] I. Akermi, Heinecke *et al.*, "Transformer based natural language generation for question-answering," in *International Conference on Natural Language Generation*, 2020, pp. 349–359.

[43] A. Ramesh, M. Pavlov *et al.*, "Zero-shot text-to-image generation," in *ICML*. PMLR, 2021, pp. 8821–8831.

[44] Z. Zhang, L. Han *et al.*, "Sine: Single image editing with text-to-image diffusion models," in *CVPR*, 2023, pp. 6027–6037.

[45] H. Lu, H. Tunanyan *et al.*, "Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models To Learn Any Unseen Style," in *CVPR*, 2023, pp. 14 267–14 276.

[46] J. Xu, X. Wang *et al.*, "Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models," in *CVPR*, 2023, pp. 20 908–20 918.

[47] X. Xu, J. Guo *et al.*, "Prompt-Free Diffusion: Taking" Text" out of Text-to-Image Diffusion Models," *arXiv preprint arXiv:2305.16223*, 2023.

[48] J. Xu, S. Liu *et al.*, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *CVPR*, 2023, pp. 2955–2966.

[49] C. Mou, X. Wang *et al.*, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.

[50] A. Nichol, P. Dhariwal, A. Ramesh *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[51] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[52] C. Saharia, W. Chan, S. Saxena *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[53] W. Hong, M. Ding, W. Zheng *et al.*, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[54] U. Singer, A. Polyak *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[55] R. Villegas, M. Babaeizadeh, P.-J. Kindermans *et al.*, "Phenaki: Variable length video generation from open domain textual description," *arXiv preprint arXiv:2210.02399*, 2022.

[56] W. Chen, J. Wu *et al.*, "Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models," *arXiv preprint arXiv:2305.13840*, 2023.

[57] A. Ramesh, P. Dhariwal *et al.*, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[58] S. Gu, Chen *et al.*, "Vector Quantized Diffusion Model for Text-to-Image Synthesis," in *CVPR*, 2022, pp. 10 696–10 706.

[59] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *NeurIPS*, vol. 30, 2017.

[60] H. Chefer, Y. Alaluf *et al.*, "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models," *SIGGRAPH 2023*, 2023.

[61] R. Gal, Y. Alaluf *et al.*, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[62] M. Petrovich, M. Black, and G. Varol, "TEMOS: Generating diverse human motions from textual descriptions," pp. 480–497, 2022.

[63] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *ECCV*. Springer, 2022, pp. 358–374.

[64] K. Youwang, K. Ji-Yeon, and T.-H. Oh, "Clip-actor: Text-driven recommendation and stylization for animating human meshes," in *ECCV*, 2022, pp. 173–191.

[65] F. Hong, M. Zhang, Z. Liu *et al.*, "AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars," *ACM Trans. on Graphics*, vol. 41, no. 4, pp. 1–19, 2022.

[66] I. Han, S. Yang, T. Kwon, and J. C. Ye, "Highly personalized text embedding for image manipulation by stable diffusion," *arXiv preprint arXiv:2303.08767*, 2023.

[67] S. Kim, C. Kim, and J. H. Park, "Human-like arm motion generation for humanoid robots using motion capture database," in *International Conference on Intelligent Robots and Systems*, 2006, pp. 3486–3491.

[68] C. Ott, D. Lee, and Y. Nakamura, "Motion capture based human motion recognition and imitation by direct marker control," in *IEEE International Conference on Humanoid Robots*, 2008, pp. 399–405.

[69] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "AI choreographer: Music conditioned 3D dance generation with aist++," in *ICCV*, 2021, pp. 13 401–13 412.

[70] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3D human motion synthesis with transformer vae," in *ICCV*, 2021, pp. 10 985–10 995.

[71] C. Ahuja and L. Morency, "Language2Pose: Natural Language Grounded Pose Forecasting," in *International Conference on 3D Vision (3DV)*, sep 2019, pp. 719–728.

[72] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[73] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt, "Mofusion: A Framework for Denoising-Diffusion-Based Motion Synthesis," in *CVPR*, 2023.

[74] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What comprises a good talking-head video generation?: A survey and benchmark," *ArXiv*, vol. abs/2005.03201, 2020.

[75] H. Zhou, Y. Liu, Z. Liu *et al.*, "Talking face generation by adversarially disentangled audio-visual representation," in *AAAI*, vol. 33, no. 01, 2019, pp. 9299–9306.

[76] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *arXiv preprint arXiv:1801.01442*, 2017.

[77] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning Lip Sync from Audio," *ACM Trans. on Graphics*, 2017.

[78] S. Si, J. Wang, X. Qu, N. Cheng, W. Wei, X. Zhu, and J. Xiao, "Speech2Video: Cross-Modal Distillation for Speech to Video Generation," in *INTERSPEECH*, 2021.

[79] F.-T. Hong, L. Zhang, L. Shen, and D. Xu, "Depth-aware generative adversarial network for talking head video generation," in *CVPR*, 2022, pp. 3397–3406.

[80] Y. Zhou, X. Han, E. Shechtman, J. Echevarria *et al.*, "MakeItTalk: Speaker-Aware Talking-Head Animation," *ACM Trans. on Graphics*, vol. 39, no. 6, pp. 1–15, 2020.

[81] M. Stypułkowski, K. Vougioukas *et al.*, "Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation," in *https://arxiv.org/abs/2301.03396*, 2023.

[82] K. Vougioukas *et al.*, "Realistic speech-driven facial animation with gans," *IJCV*, vol. 128, pp. 1398–1413, 2020.

[83] Z. Zhang, L. Li *et al.*, "Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

[84] M. Lee, K. Lee, and J. Park, "Music Similarity-Based Approach to Generating Dance Motion Sequence," *Multimedia Tools Appl.*, vol. 62, no. 3, p. 895–912, 2013.

[85] R. Fan, S. Xu *et al.*, "Example-Based Automatic Music-Driven Conventional Dance Motion Synthesis," *IEEE TVCG*, vol. 18, no. 3, pp. 501–515, 2012.

[86] F. Ofli, E. Erzin *et al.*, "Learn2Dance: Learning Statistical Music-to-Dance Mappings for Choreography Synthesis," *IEEE TMM*, vol. 14, no. 3, pp. 747–759, 2012.

[87] O. Alemi, J. Francoise, and P. Pasquier, "GrooveNet: Real-time music-driven dance movement generation using artificial neural networks," *networks*, vol. 8, no. 17, p. 26, 2017.

[88] H.-K. Kao and L. Su, "Temporally guided music-to-body-movement generation," in *ACM International Conference on Multimedia*, 2020, pp. 147–155.

[89] X. Ren, H. Li, Z. Huang, and Q. Chen, "Self-supervised dance video synthesis conditioned on music," in *ACM International Conference on Multimedia*, 2020, p. 46–54.

[90] L. Siyao, W. Yu *et al.*, "Bailando: 3D dance generation by actor-critic gpt with choreographic memory," in *CVPR*, 2022, pp. 11 050–11 059.

[91] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Trans. Graph.*, 2023.

[92] J. Li, Y. Yin, H. Chu, Y. Zhou, T. Wang, S. Fidler, and H. Li, "Learning to generate diverse dance motions with transformer," *CoRR*, 2020.

[93] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *ICLR*, 2021.

[94] P. Isola, J.-Y. Zhu *et al.*, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[95] Y. Pang, J. Lin *et al.*, "Image-to-image translation: Methods and applications," *IEEE TMM*, vol. 24, pp. 3859–3881, 2021.

[96] J. Han, M. Shoeiby *et al.*, "Dual contrastive learning for unsupervised image-to-image translation," in *CVPR*, 2021, pp. 746–755.

[97] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu *et al.*, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018, pp. 8798–8807.

[98] T. Park, M.-Y. Liu *et al.*, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, 2019, pp. 2337–2346.

[99] T. Kim, M. Cha *et al.*, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017, pp. 1857–1865.

[100] Y.-J. Chen, S.-I. Cheng *et al.*, "Vector Quantized Image-to-Image Translation," in *ECCV*, 2022.

[101] C. Saharia, W. Chan, H. Chang *et al.*, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH*, 2022, pp. 1–10.

[102] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models," 2021.

[103] B. Li, K. Xue *et al.*, "BBDM: Image-to-Image Translation With Brownian Bridge Diffusion Models," in *CVPR*, 2023, pp. 1952–1961.

[104] T. Chen, M.-M. Cheng *et al.*, "Sketch2Photo: Internet Image Montage," *ACM Trans. Graph.*, vol. 28, no. 5, p. 1–10, 2009.

[105] M. Eitz, R. Richter *et al.*, "Photosketcher: Interactive Sketch-Based Image Synthesis," *IEEE Computer Graphics and Applications*, vol. 31, pp. 56–66, 2011.

[106] W. Chen and J. Hays, "SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis," in *CVPR*, June 2018.

[107] Y. Li, X. Chen *et al.*, "LinesToFacePhoto: Face Photo Generation From Lines With Conditional Self-Attention Generative Adversarial Networks," *ACM International Conference on Multimedia*, 2019.

[108] A. K. Bhunia, P. N. Chowdhury, A. Sain *et al.*, "More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval," in *CVPR*, 2021, pp. 4247–4256.

[109] S. Koley, A. K. Bhunia *et al.*, "Picture That Sketch: Photorealistic Image Generation From Abstract Sketches," in *CVPR*, 2023, pp. 6850–6861.

[110] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J Acoust Soc Am*, vol. 120, pp. 2421–2424, 2006.

[111] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE TAC*, vol. 5, pp. 377–390, 10 2014.

[112] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *ACL*, 2018.

[113] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[114] Y. Tian, D. Li, and C. Xu, "Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing," in *ECCV*, 2020.

[115] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *ICCV*, 2019, pp. 5442–5451.

[116] C. Guo, X. Zuo, S. Wang *et al.*, "Action2motion," in *ACM International Conference on Multimedia*. ACM, 2020.

[117] A. R. Punnakkal, A. Chandrasekaran *et al.*, "BABEL: Bodies, action and behavior with english labels," in *CVPR*, 2021, pp. 722–731.

[118] C. Guo, S. Zou *et al.*, "Generating Diverse and Natural 3D Human Motions From Text," in *CVPR*, 2022, pp. 5152–5161.

[119] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big data*, vol. 4, no. 4, pp. 236–252, 2016.

[120] G. Lee, Z. Deng *et al.*, "Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis," in *ICCV*, 2019, pp. 763–772.

[121] Y. Ferstl and R. Mcdonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," *International Conference on Intelligent Virtual Agents*, 2018.

[122] T. Tang, J. Jia, and H. Mao, "Dance with Melody: An LSTM-Autoencoder Approach to Music-Oriented Dance Synthesis," in *ACM International Conference on Multimedia*, 2018, p. 1598–1606.

[123] C. Kang, Z. Tan *et al.*, "ChoreoMaster : Choreography-Oriented Music-Driven Dance Synthesis," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, 2021.

[124] R. Li, J. Zhao *et al.*, "FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation," in *ICCV*, 2023.

[125] C. Wah, S. Branson, P. Welinder, P. Perona, and S. J. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:16119123

[126] T.-Y. Lin, M. Maire *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*. Springer International Publishing, 2014.

[127] J. Deng, W. Dong *et al.*, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[128] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," *ArXiv*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:241033103

[129] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *ArXiv*, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:7197134

[130] Y. Jiang, Z. Huang *et al.*, "Talk-to-Edit: Fine-Grained Facial Editing via Dialog," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.

[131] S. Changpinyo, P. Sharma *et al.*, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021.

[132] T. Salimans, I. Goodfellow, W. Zaremba *et al.*, "Improved techniques for training gans," in *NeurIPS*, vol. 29, 2016.

[133] M. Heusel, H. Ramsauer *et al.*, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *NIPS*, vol. 30, 2017.

[134] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *ArXiv*, vol. abs/1812.01717, 2018.

[135] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[136] J. A. Aslam and E. Yilmaz, "A Geometric Interpretation and Analysis of R-Precision," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005.

[137] Z. Zhou and B. Wang, "UDE: A Unified Driving Engine for Human Motion Generation," in *CVPR*, 2023.

[138] J. Hessel *et al.*, "Clipscore: A reference-free evaluation metric for image captioning," in *2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528.

[139] M. Toshpulatov, W. Lee, and S. Lee, "Talking human face generation: A survey," *Expert Systems with Applications*, vol. 219, p. 119678, 2023.

[140] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies," *ACM Trans. Graph.*, vol. 35, 2016.

[141] Z. W. Xingqian Xu *et al.*, "Versatile Diffusion: Text, Images and Variations All in One Diffusion Model," in *ICCV*, 2023.

[142] J. Xing, M. Xia, Y. Zhang *et al.*, "CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior," *CVPR*, 2023.

[143] F. Yin, Y. Zhang, X. Cun *et al.*, "StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN," *ECCV2022*, 2022.

[144] Z. Wang, H. Zheng *et al.*, "Diffusion-GAN: Training GANs with Diffusion," *arXiv preprint arXiv:2206.02262*, 2022.

Fatemeh Nazarieh is currently a PhD student at the School of Computer Science and Electronic Engineering, the University of Surrey. Her research is mainly focused on cross-modality content generation, particularly in the area of audio-to-talking face generation. She has previously published papers in venues such as Wiley and ACM. Her research aims to bridge the gap between audio and visual data, enhancing the realism and synchronization of digital representations in multimedia applications.

Zhenhua Feng (S'13-M'16-SM'22) received the Ph.D. degree from the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K. in 2016. He is currently a Senior Lecturer in Machine Learning and Computer Vision at the School of Computer Science and Electronic Engineering. His research interests include pattern recognition, machine learning, and computer vision.

He has published more than 90 scientific papers in top-tier conferences and journals, such as TPAMI, IJCV, CVPR, ICCV, IJCAI, AAAI, ACL, TIP, TNNLS, TCYB, TIFS, etc. He currently serves as the Associate Editor for IEEE TNNLS and Complex & Intelligent Systems. He also served as the Programme Chair of BMVC 2022, Area Chair of BMVC 2021/22/23 & CVMP 2022/23, and Senior Programme Committee Member of IJCAI 2021. He has received the 2017 European Biometrics Industry Award from the European Association for Biometrics (EAB), and the AMDO 2018 Best Paper Award for Commercial Applications.

Muhammad Awais is currently a senior lecturer jointly at Centre for Vision Speech and Signal Processing (CVSSP), and Surrey Institute for People-centred AI (SI-PAI), University of Surrey. His research interests include self-supervised learning, computer vision, medical image analysis, multimodal learning, pattern recognition, machine learning and deep learning. He received his BSc degree in mathematics and physics in 2001, the BSc degree in computer engineering, in 2005, the MSc degree in signal processing and machine intelligence and the PhD degree in machine learning, in 2008 and 2011.

Wenwu Wang is currently a Professor of signal processing and machine learning, and the Co-Director of the Machine Audition Lab, Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K. He is also an AI Fellow with the Surrey Institute for People Centred Artificial Intelligence. His research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has co-authored more than 350 papers in these areas. He is involved as Principal or Co-Investigator in more than 30 research projects, funded by U.K. and EU research councils, and industry which include BBC, NPL, Samsung, Tencent, Huawei, Saab, Atlas, and Kaon. He is the elected Chair of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee, the elected Vice Chair of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, a Board Member of IEEE SPS Technical Directions Board. He is an Associate Editor for the IEEE/ACM Transactions on Audio Speech and Language Processing, an Associate Editor for (Nature) Scientific Report, and Specialty Editor in Chief of Frontier in Signal Processing. He was the Senior Area Editor during 2019–2023, and an Associate Editor during 2014–2018, for IEEE Transactions on Signal Processing. He is an invited Keynote or Plenary Speaker on more than 20 international conferences and workshops, and a Member of the technical program committee for more than 100 international conferences or workshops.

Josef Kittler (M'74-LM'12) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook Pattern Recognition: A Statistical Approach and over 700 scientific papers. His publications have been cited around 70,000 times (Google Scholar).

He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, Pattern Analysis and Applications. He also served as a member of the Editorial Board of IEEE Transactions on Pattern Analysis and Machine Intelligence during 1982-1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982-2005, President of the IAPR during 1994-1996.