# Multi-modal Generative AI: Multi-modal LLMs, Diffusions and the Unification

Xin Wang, *Member, IEEE*, Yuwei Zhou, Bin Huang, Hong Chen, and Wenwu Zhu, *Fellow, IEEE*

***Abstract*—Multi-modal generative AI (Artificial Intelligence) has attracted increasing attention from both academia and industry. Particularly, two dominant families of techniques have emerged: i) Multi-modal large language models (LLMs) demonstrate impressive ability for *multi-modal understanding*; and ii) Diffusion models exhibit remarkable multi-modal powers in terms of *multi-modal generation*. Therefore, this paper provides a comprehensive overview of multi-modal generative AI, including multi-modal LLMs, diffusions, and the unification for understanding and generation. To lay a solid foundation for unified models, we first provide a detailed review of both multi-modal LLMs and diffusion models, respectively, including their probabilistic modeling procedure, multi-modal architecture design, and advanced applications to image/video LLMs as well as text-to-image/video generation. Furthermore, we explore the emerging efforts toward unified models for understanding and generation. To achieve the unification of understanding and generation, we investigate key designs including autoregressive-based and diffusion-based modeling, as well as dense and Mixture-of-Experts (MoE) architectures. We then introduce several strategies for unified models, analyzing their potential advantages and disadvantages. In addition, we summarize the common datasets widely used for multi-modal generative AI pretraining. Last but not least, we present several challenging future research directions that may contribute to the ongoing advancement of multi-modal generative AI.**

***Index Terms*—Multi-modal Generative AI, Multi-modal Large Language Model, Diffusion Model, Unified Understanding and Generation**

## I. Introduction

Multi-modal generative AI (Artificial Intelligence) has received increasing attention recently with the advent of (multi-modal) large language models (LLMs) and diffusion models. Two typical models of multi-modal generative AI are GPT-4V [1] and Sora [2] from OpenAI, which have produced great impacts on both academia and industry. To compare GPT-4V and Sora in terms of functionality, GPT-4V targets multi-modal understanding, and Sora aims at visual generation — GPT-4V enables the LLM to understand visual input via generating relevant texts, while Sora serves as a text-to-video generation model which outputs visual signals given textual input. To make comparisons in terms of probabilistic modeling, GPT-4V is a multi-modal LLM with autoregressive probabilistic modeling, while Sora is a multi-modal video generation model with diffusion denoising modeling.

As such, there naturally arises a question: "Is it possible to establish a unified multi-modal generative model for simultaneous understanding and generation?" And if the answer is *yes*, what would such a model be, either similar to multi-modal LLM or diffusion, or in a new form? To capture the relations among different modalities, is it a good idea to adopt an early-fusion strategy (such as Chameleon [3]), or just straightforwardly align a pretrained visual model with a language model (such as LLAVA [4])? To further unify understanding and generation, is it sufficient to employ Mixture of Experts (MoE) strategies or only use a dense model?

To answer these questions, we conduct deep and comprehensive discussions of multi-modal generative AI in this paper, whose overall organization is illustrated in Fig. 1. Specifically, we first present a systematic review of existing works on multi-modal LLM (Sec. II) and multi-modal diffusion (Sec. III), covering mathematical preliminaries, model architectures, fusion strategies, recent advances, and applications. Then we present our insights on unified models for simultaneous understanding and generation in Sec. IV. Besides, we further summarize video/visual-language datasets for multi-modal generative AI pretraining in Sec. V. Last, we provide future directions that deserve further investigation for multi-modal generative AI.

In this paper, our scope primarily lies in multi-modal understanding, generation, and their unification. Some concepts widely studied in the field of LLMs, such as in-context learning, post-training techniques (e.g., supervised fine-tuning and reinforcement learning), sparse attention, and positional embeddings, are important but not the main focus of this survey. Readers interested in these topics are referred to related surveys such as [5], [6]. Instead, we focus on recent high-quality works adapted to the multi-modal generative setting, providing a comprehensive overview of the mechanisms that enable multi-modal understanding and generation.

We would like to point out that although several insightful surveys have been conducted on multi-modal understanding [7]–[9], visual generation [10]–[14], and both [15], [16], this work differs from them in comprehensive discussions on models for the unification of understanding and generation in addition to reviewing them separately, thus contributing to the ongoing advancement of multi-modal generative AI. We highlight recent advances, categorize existing approaches, introduce related datasets, and share insights for future directions. In summary, we make the following contributions.

Xin Wang, Yuwei Zhou, Bin Huang, Hong Chen, and Wenwu Zhu are with the Department of Computer Science, Beijing Information Science and Technology National Research Center, Tsinghua University, Beijing 100084, China. (E-mail: {xin_wang, wwzhu}@tsinghua.edu.cn), {zhouyw21, huangb23, h-chen20}@mails.tsinghua.edu.cn.

Corresponding author: Wenwu Zhu

This work was supported by the National Natural Science Foundation of China No. 62222209, Beijing National Research Center for Information Science and Technology under Grant No. BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

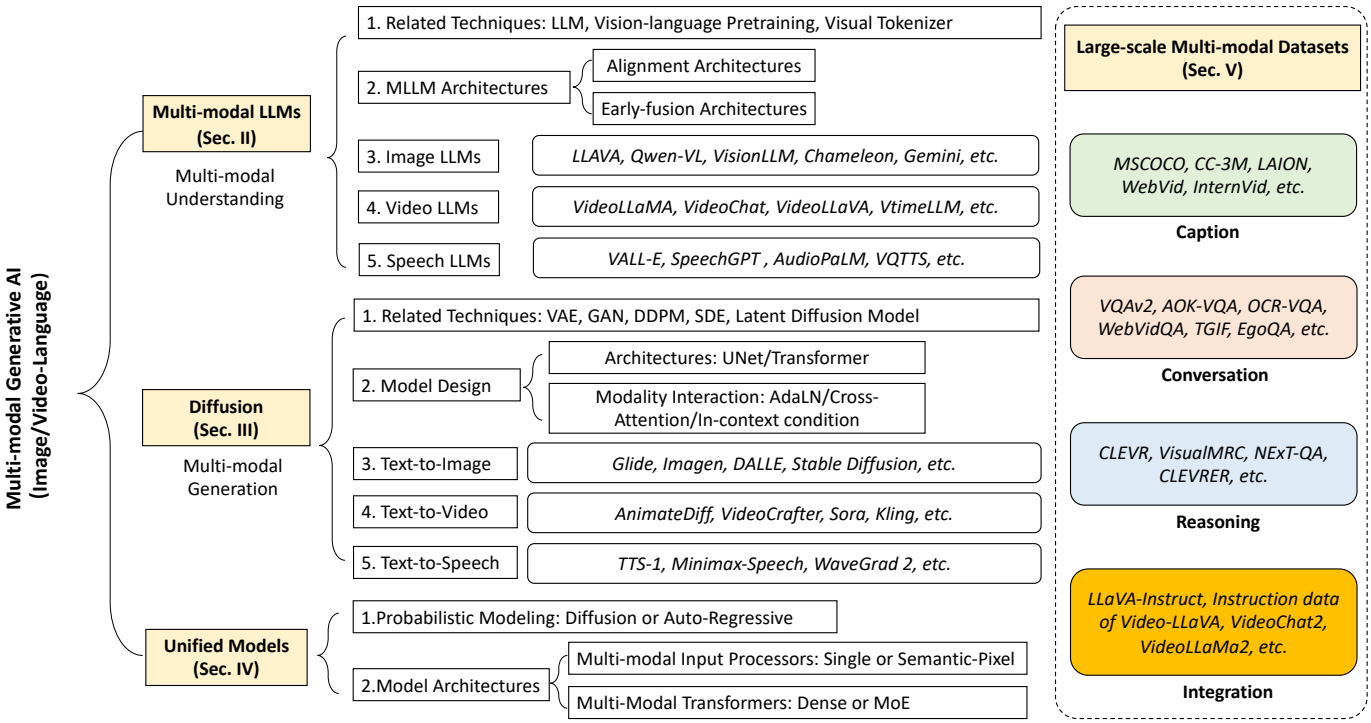


Fig. 1. The overall organization of this paper.

- We comprehensively overview multi-modal generative AI, covering multi-modal LLMs for multi-modal understanding and diffusion models for visual generation.
- We propose a structured taxonomy of unified models for multi-modal understanding and generation, and provide thorough discussions on them.
- We share our insights on promising future directions to highlight the trending research for advances in multi-modal generative AI.

## II. Multi-modal LLM for Understanding

Multi-modal LLMs have recently become dominant in the field of understanding. In this section, we will review the literature on the multi-modal LLMs.

### A. Preliminaries

We first introduce some preliminaries involving the LLM, vision-language pretraining, and visual tokenizers.

*1) LLM Autoregressive Probabilistic Modeling:* The core component of multi-modal LLMs is the LLM, which receives the multi-modal input, including the user's instructions, questions, and visual information, and then outputs the answers to the user in a text-generation form. The LLM is basically an autoregressive model that tries to predict the next word based on all the previous words, as shown in Eq. (1).

$$p(w) = \prod_{i=1}^{n} p_{\theta_L}(w_i|w_{<i}), \tag{1}$$

where $\theta_L$ denotes the parameters of the LLM, which is generally composed of several layers of transformers [17]. Note that LLM can only receive the text tokens as its input. The next important problem for multi-modal LLM is how to enable LLM to understand the visual information. To tackle the problem, most existing works [4], [18], [19] try to align the LLM with the visual encoders from vision-language pretraining tasks, such as CLIP [20]. More recently, there have been some attempts [3] to directly transform the images into discrete visual tokens so that the text and visual tokens can be tackled by the autoregressive LLM together. Next, we will introduce preliminaries about vision-language pretraining and visual tokenizers.

*2) Vision-Language Pretraining:* Vision-language pretraining (VLP) aims to learn aligned representations of images and texts by leveraging large-scale image-text pairs. One of the most influential VLP models is CLIP [20], which learns a joint embedding space where semantically related images and texts are mapped close to each other.

CLIP consists of two separate encoders: a visual encoder (typically a Vision Transformer [21] or ResNet [22]) and a text encoder (usually a Transformer). Given a batch of image-text pairs, CLIP is trained with a contrastive loss that encourages the embeddings of matched image-text pairs to be close while pushing apart the embeddings of mismatched pairs.

The pretrained CLIP model has been widely used in multi-modal LLMs to inject visual understanding into LLMs. Typically, visual features extracted by the CLIP image encoder are projected into the input space of LLM through a learned adapter or alignment module [4]. This allows LLMs to reason over both linguistic and visual information in a unified manner.

*3) Visual Tokenizer:* Inspired by language models where each word is tokenized by a discrete tokenizer, a series of works also transform images into discrete tokens. Typical

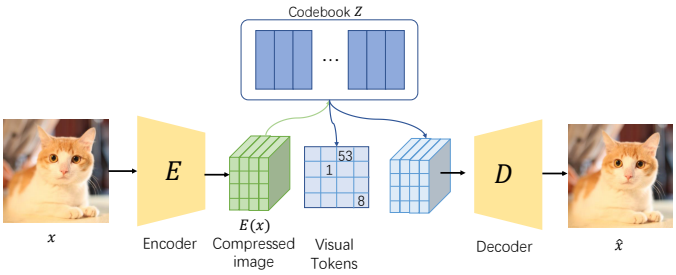


Fig. 2. Illustration for the framework of the visual tokenizers.

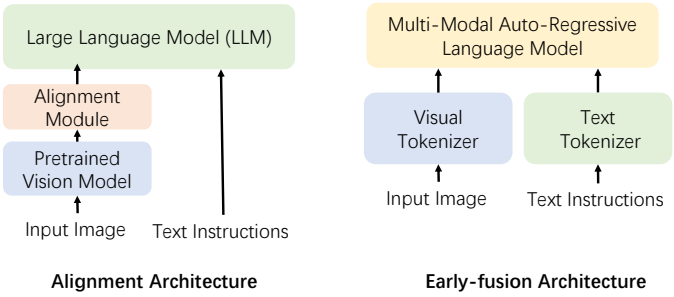


Fig. 3. Two branches of multi-modal LLM architectures, including (i) the alignment architecture by aligning pretraining vision models with LLM and (ii) the early-fusion architecture which receives mixed visual and text tokens and relies on autoregressive modeling for multi-modal understanding.

visual tokenizers include the VQ-VAEs [23], [24] and VQ-GANs [25], [26], whose overall framework is shown in Fig. 2. We will begin our discussion with VQ-VAE. Basically, VQ-VAE works as an auto-encoder with an encoder $E(\cdot)$ and a decoder $D(\cdot)$. Given an image $x$, VQ-VAE first encodes it with an encoder $E(\cdot)$ into a lower-dimensional continuous vector $E(x)$. Then, the continuous vector is discretized using a codebook $Z = \{z_k\}_{k=1}^{K}$. The codebook functions similarly to a word embedding table in NLP, where $K$ corresponds to the vocabulary size, and each $z_k \in \mathbb{R}^{n_c}$ represents a visual prototype analogous to a word embedding. With the encoded vector $E(x)$ and the codebook $Z$, we obtain a discrete representation $z_q$ of the image by finding the nearest neighbor of $E(x)$ in $Z$ and use it to reconstruct the image with the decoder: $\hat{x} = D(z_q)$. This provides a way to convert between images and discrete tokens.

Compared to VQ-VAEs, VQGAN [25], [26] utilizes a GAN perceptual loss to replace the L2 reconstruction loss, which helps to learn a rich codebook. We use a simple example to illustrate the tokenization process. If we have an input image of size $H \times W \times 3$, after the encoder $E$, we obtain a lower-dimension vector $E(x)$ of size $h \times w \times n_c$, where $h < H$, $w < W$, and $n_c$ denote the dimensions of the code. This means that we can obtain $h \times w$ vectors of dimension $n_c$, and for each vector we will find its nearest neighbor in the code book for discretization so that we will finally obtain a discrete sequence of length $h \times w$ to represent the image.

**Remark.** On the one hand, VQGAN and VQ-VAE can be used as visual tokenizers to transform an image into discrete tokens, which enables it to be received by LLMs for visual understanding. On the other hand, they can be used to compress an image into a lower-dimensional space, which motivates the well-known latent diffusion model (LDM) [27].

### B. Multi-modal LLM Architectures

We categorize existing multi-modal LLM architectures into two branches, the alignment architectures and the early-fusion architectures, as shown in Fig. 3. Most existing works [4], [18], [19] adopt the alignment architecture, which aims to align the vision model from the vision-language pretraining with the pretrained LLM. This branch of models relies on the vision-language pretraining to understand the visual input. After obtaining the embedding of the image, an alignment module such as a projector [4] or Q-Former [28] is used to align the image embedding with the LLM space. To train the alignment module, some text-image or text-video pairs are required to input the model. A typical way to align is to make the LLM output the caption of an image given an image embedding. In contrast, as shown on the right of Fig. 3, the early-fusion architectures [3], [29] do not rely on a pretrained vision model to obtain the semantics of the input image. Instead, similar to NLP, where each word is mapped to a token, the early-fusion architecture maps each visual input into visual tokens through a visual tokenizer. Then, a multi-modal autoregressive language model will receive the mixed text and visual tokens and output the user's desired answers.

Next, with the overall architecture in mind, we will introduce recent advances in image LLMs and video LLMs.

### C. Image LLM

We will follow the multi-modal LLM architectures section and elaborate on the latest advancement of image LLM.

*1) Alignment-Architecture Image LLM:* This architecture treats the image input as an additional extension. The vision encoders are usually frozen and the alignment modules and LLM are tuned based on various strategies to align the multi-modal content and instructions.

**a) Vision Encoder** is a module that extracts crucial information from images. Common generic vision encoders include ResNet [30], the CLIP-ViT encoder [20], and ImageBind [31]. ResNet and CLIP are pretrained on image-text modalities, while ImageBind aligns embeddings from six modalities into one shared space, enabling vision encoders to capture richer information.

**b) Alignment Module**, also named projector, adapter, etc., aims to mitigate the gap between image features and lexical word tokens and further fuse two modalities. LLaVA [32] adopts a simple but effective linear projection to convert image features into word token embedding space and then it concatenates image tokens and word tokens. Such alignment only involves image transformation, limiting interaction with texts, and is not flexible in the visual token number. Resampler [33] technique maps varying-size features to a fixed number of tokens. BLIP-2 [28] and MiniGPT-4 [34] employ Q-former [28] before linear projections to reduce tokens. Q-former incorporates text semantics and models the interaction between image features and text inputs with learnable queries to enhance the most useful visual content for LLM. Some

works focus on preserving locality during projection, such as Honeybee [35], which introduces a locality-enhanced projector to maintain spatial structure. Others prioritize efficiency, such as TokenPacker [36], which adopts a coarse-to-fine strategy to compress visual tokens while retaining important details.

*2) Early-fusion Architecture Image LLM:* The alignment architecture utilizes the power of off-the-shelf LLM and requires lower computations, but pretrained vision encoders would have information loss and be infected by inductive biases because of the gap between limited pretraining tasks and real demands for image LLM, such as supporting flexible resolution. Therefore, as shown in Fig. 3, another line of work aims to train a multi-modal LLM from scratch, where both images and text words are converted into a series of tokens.

Pioneer work Fuyu [37] adopts linear projections on image patches in spatial order and trains a transformer decoder taking the visual and word token sequence as input. Despite limited performance, it reveals a new technical fashion. Google follows this fashion, whose Gemini [29] processes the interleaved image and other modalities from the beginning. Chameleon [3] trains an image tokenizer that encodes a 512x512 image into 1024 discrete tokens from a codebook of size 8192. Early-fusion Architecture requires more computation and is more difficult to converge, leaving challenges for future exploration.

*3) Challenges in Image LLM:* **(i) Fine-grained visual concept understanding**, where more tokens help encode more detailed information at the cost of causing redundant computation. Chat-UniVi [38] proposes dynamic visual tokens to allocate more computations on important details. An important part of fine-grained understanding is the spatial awareness of object concepts. AnyRef [39] applies RoIAlign to encode regions and designs a segment encoder-decoder to learn segmentation from the image LLM's token outputs, which is similar to OMG-LLaVA [40], who generates pixel- and object-centric visual tokens before projections and decodes segmentation tokens from LLM's output by OMG-Seg. Different from segmentation supervision, VisionLLM [41] and Virtron [42] use text supervision such as bounding and polygon descriptions by flexible instruction tuning. Fine granularity modeling offers some explanations for LLM. **(ii) Hallucination** involves errors in objects, attributes, and relations in the forms of judgment or description [43]. Some works [44] try to reduce biases in training data, while some mitigate hallucination by improving model characteristics such as vision encoders [45] or fusion mechanisms [46]. Human feedbacks [47] also play an important role in reducing hallucination.

**Remark.** Currently, the alignment architecture still outperforms the early-fusion architecture in multi-modal understanding, e.g., with comparable parameters, the early-fusion architecture Emu3 [48] achieves 75.1 score on VQAv2 [49] benchmark and 58.5 score on MMBench [50] benchmark, while the early-fusion architecture LLAVA-1.6 achieves 86.8 and 67.4 score, respectively. The advantages and disadvantages of the two architectures are as follows: (i) The advantage lies in the capability of utilizing the pretrained knowledge from the vision encoder and LLM. The vision-language pretraining enables the output of the vision encoder to contain semantic meanings. Only the alignment module needs to be trained, which makes this paradigm resource-friendly. (Sometimes other modules are also learnable for better performance.) However, its ability is also limited by the pretrained vision encoder and LLM, e.g., the pretrained CLIP vision encoder often struggles with multiple objects, making the multi-modal LLMs based on CLIP inherit the limitation. (ii) The disadvantage comes from the fact that the early-fusion architecture may have a higher potential, because all its parameters are trained from scratch. However, training from scratch makes the early-fusion architecture face two challenges: (a) a good visual tokenizer needs to be trained, and (b) more resources will be needed to train the multi-modal autoregressive model. First, since the visual tokenization process involves compression and discretization, there inevitably exists visual information loss. How to train a tokenizer that contains rich visual information still remains a challenging problem. Second, the visual tokenizers are generally trained with the image reconstruction objective, which in essence belongs to a pixel-level task instead of a semantic-level task. This training strategy requires that the downstream multi-modal LLMs should have an additional ability to learn semantic meanings from the pixel-level information, compared to the original LLMs, which are only expected to understand semantic tokens. Therefore, multi-modal LLMs tend to require more data for training.

### *D. Video LLM*

Following the success of Image LLMs, researchers start exploring the training of Video LLMs [51]. Typically, videos are viewed as sequences of image frames (some Video LLMs incorporate other modalities like audio or speech), so Video LLMs have a higher computational complexity. The challenge of collecting high-quality video datasets further complicates the training process, making early fusion architectures computationally exhaustive. As a result, almost all the existing Video LLMs adopt the alignment architectures.

*1) Alignment-Architecture Video LLM:* The video LLM architecture is similar to that of Image LLMs with alignment architectures. By sampling a fixed number of frames or using a fixed frames-per-second (FPS) rate, videos are reduced to a limited set of images. The visual embeddings of each image are then extracted using a visual encoder. These features are sequentially concatenated in the order of the frames and connected to the LLM via an alignment module. In earlier works, VideoChat [52] utilizes a Q-former structure as the alignment module, while VideoLLaMA [53] introduces an audio encoder and an audio Q-former to handle audio signals. Video-ChatGPT [54] takes a different approach by average-pooling each frame's patch embeddings along the spatial and temporal dimensions before using a linear layer as the alignment module. Training Video LLMs also follow an "alignment then instruction tuning" strategy. While additional GPT-annotated or human-annotated video datasets are collected, image datasets can also be leveraged by treating images as single-frame videos.

Recent successful efforts focus on improving performance by refining the alignment module and scaling up the model and dataset sizes. For instance, VideoLLaMA2 [55] improves the

alignment module to model the connections across temporal and spatial dimensions. It also gathers datasets for tasks such as captioning, classification, and question answering. Qwen2.5-VL [56] and InternVL3 [57] leverage diverse training data, including images, videos, and interleaved image–text pairs, to build powerful vision-language models.

*2) Challenges and Limitations in Video LLM:* Compared to Image LLMs, Video LLMs face two unique challenges. The first challenge is understanding videos at a finer granularity, specifically the comprehension of video segments and the relationships between these segments. The second challenge is understanding long-form videos, such as movies, within the limited context length of LLMs.

For segment-level video understanding, VTimeLLM [18] transforms the temporal video grounding and dense video captioning tasks into a sequence-to-sequence format. After alignment training, it introduces an additional boundary perception training, leveraging large-scale multi-event video-text data to enhance awareness of event boundaries and timestamps. Finally, it incorporates temporal reasoning data during instruction tuning. Some approaches [58], [59] adopt training-free methods, where sampled frames are individually captioned, and each frame's timestamp and caption are input into an LLM via carefully crafted prompts, allowing the LLM's powerful reasoning capabilities to comprehend each segment.

For long-form videos, traditional Video LLMs struggle with input limitations. For example, a Q-former in BLIP-2 encodes an image into 32 tokens; sampling 256 frames results in 8K tokens, which reaches the maximum context length of most LLMs. However, this represents less than 5 minutes of video at a sampling rate of 1 FPS. Therefore, more efficient representations are necessary for processing long-form videos like movies. MovieChat [60] introduces a memory consolidation mechanism that merges similar image tokens once the token limit is reached. LWM [61] and LongVA [62] handle long video inputs by using LLMs with larger context lengths and more efficient attention mechanisms. Some methods [18], [63] reduce the number of tokens per frame, representing each frame with only 1 or 2 tokens on average. Other approaches [64] convert long-form videos into text corpus using image captioning and employ LLMs as agents to search for specific answers within the text corpus.

**Remark.** Despite the advancements in Video LLMs, nearly all existing models rely on sampling frames and encoding them individually through image encoders. This approach may be favored due to several reasons: image encoders are less computationally intensive compared to video encoders, they offer better alignment with textual data, and they facilitate unification with Image LLMs. However, this methodology comes with a significant limitation. Specifically, the process of sampling frames can lead to the complete loss of information that occurs between sampled frames. As a result, these models fail to capture the continuous motion and trajectories of objects, which are essential for understanding dynamic scenes and activities within a video.

### *E. Speech LLM*

Similar to Image LLMs, the architecture of Speech LLMs can generally be categorized into two types: alignment-based architectures and early-fusion architectures [65].

*1) Alignment-Architecture Speech LLM:* This architecture first extracts information from audio with pre-trained or fine-tuned audio encoder and produces audio embedding.

**a) Audio Encoder** transforms raw waveforms into time–frequency representations using conventional signal processing techniques. The most commonly used audio encoders are Whisper [66] and Conformer [67]. Whisper is an automatic speech recognition (ASR) model with an encoder–decoder Transformer architecture, similar to sequence-to-sequence models in natural language processing. It is trained on 680,000 hours of multilingual, multitask supervised data collected from the web, covering speech recognition, speech translation, and language identification. Conformer (Convolution-augmented Transformer) combines convolutional neural networks (CNNs) with Transformer blocks, effectively capturing both local and global dependencies in speech signals. Other widely adopted encoders include WavLM [68], a self-supervised speech representation model built on the HuBERT [69] framework, with improvements in pretraining objectives and data diversity.

**b) Alignment Module** also referred to as a projector, connector, or adapter, maps audio embeddings into the text embedding space, enabling them to be processed by the LLM decoder for downstream understanding tasks. Several types of alignment modules have been proposed. One common approach is a multi-layer perceptron (MLP), which performs a straightforward projection. Another is the Q-Former, which introduces trainable query tokens that attend to audio features and produce fixed-length embeddings compatible with the LLM input space. A third approach is cross-attention, which allows bidirectional interactions between audio and text features, facilitating richer multimodal integration.

*2) Early-fusion Architecture Speech LLM:* This type of Speech LLMs is inspired by visual tokenizers and adopts a similar approach for audio. In this framework, raw audio is converted into a sequence of discrete tokens that capture the acoustic content and can often be decoded back into high-quality audio. The generation of discrete tokens relies on vector quantization (VQ). Building on VQ-VAE [70], which introduced the idea of encoding continuous audio features into symbolic representations via a learned codebook, modern approaches include self-supervised pre-trained audio tokenizers such as HuBERT [69] and neural codec models such as EnCodec [71]. Several representative works fall under this branch of Speech LLMs. VALL-E [72] leverages EnCodec tokens to achieve zero-shot speech synthesis. SpeechGPT [73] is trained on paired unit-text data, where spoken audio is represented as discrete speech units. AudioPaLM [74] integrates wav2vec-style audio tokenization with language modeling to improve multimodal speech understanding.

Now we have discussed the multi-modal LLM for understanding. Next, we will discuss another important topic of multi-modal generative AI, i.e., multi-modal diffusion models for generation.

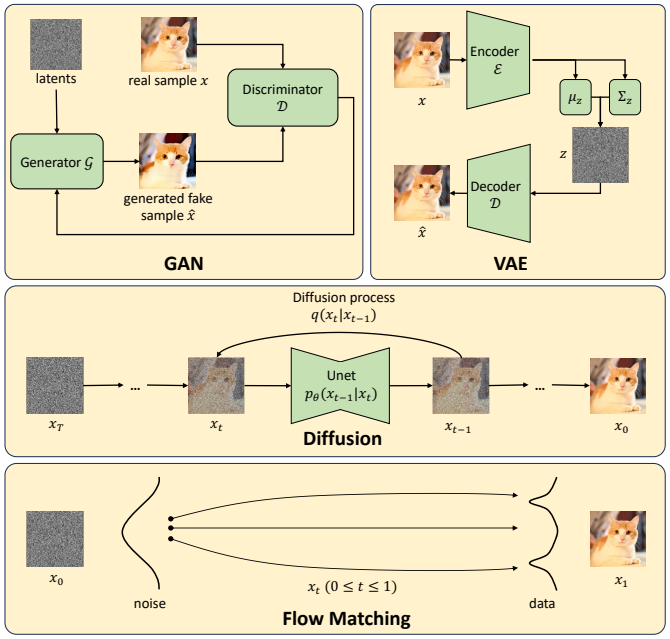


Fig. 4. Comparison among GAN, VAE, diffusion, and flow matching models.

## III. Multi-modal Diffusion for Generation

Diffusion models have been one of the most successful generative models in visual generation given texts and are widely used in multi-modal generation tasks. We present the famous latent diffusion model [27], and discuss several advanced diffusion-based text-to-image and text-to-video models.

### A. Preliminaries

We will first introduce some preliminaries, including traditional generative models, i.e., generative adversarial networks (GANs) and Variational AutoEncoders (VAEs). We then introduce diffusion probabilistic modeling and present a comparison among GAN, VAE, diffusion, and flow matching models, as illustrated in Fig. 4.

*1) Generative Adversarial Networks:* The generative adversarial network (GAN) [75] is one of the earliest neural architectures designed to generate visual content such as images [76] and videos [77]. The main idea of GANs involves two networks: a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. Specifically, $\mathcal{G}$ aims to generate visual content from a noise vector $z$, while $\mathcal{D}$ is trained to distinguish between real visual samples $x$ and generated ones $\mathcal{G}(z)$. These two networks are trained in an adversarial manner: the generator tries to produce outputs that can fool the discriminator, and the discriminator strives to accurately classify real versus fake samples. The training process forms a min-max game, where the generator learns to generate increasingly realistic samples to deceive a progressively stronger discriminator. The two networks are mutually reinforcing, so the training objective is as follows:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{x\sim p_x} \log \mathcal{D}(x) + \mathbb{E}_{z\sim p_z} \log(1 - \mathcal{D}(\mathcal{G}(z))), \quad (2)$$

where $z$ is sampled from $p_z$ that is usually a normal distribution and $x$ is a sample from the real data distribution $p_x$.

*2) Variational AutoEncoder:* Variational AutoEncoder [78] (VAE) is another typical generative model. Unlike GANs, autoencoders have an encoder-decoder architecture that uses an encoder $\mathcal{E}$ to present the visual content $x$ to a latent code $z = \mathcal{E}(x)$ and a decoder $\mathcal{D}$ to reconstruct the data $\hat{x} = \mathcal{D}(z) \approx x$. However, normal autoencoders have no constraints on the latent space, which makes them overfit the dataset easily. To solve the problem, VAEs make a regularization to the latent space and sample $z$ from a distribution $p_\theta$, typically a Gaussian distribution, where $\theta$ is the parameters of the encoder-decoder model. As the distribution $p_\theta$ is unknown, VAE utilizes a recognition model $\phi$ which serves as a variational approximation $q_\phi$ to approximate $p_\theta$ and trains them jointly:

$$\mathcal{L}(\theta, \phi; x) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)], \quad (3)$$

where $D_{KL}$ means the Kullback-Leibler divergence. $\phi$ can be formulated as a differentiable estimator using the parameterization trick. To better generate visual content, many efforts [70], [79], [80] have been made based on VAE. Sync-DRAW [79] introduces a novel architecture that combines VAE with a recurrent attention mechanism to create a unique temporally dependent sequence of frames.

Despite the successful introduction of VAEs, they still face a significant issue where the model ignores the information in the latent space and relies solely on a powerful decoder to reconstruct the data, a phenomenon known as "posterior collapse". To address this problem, the VQ-VAE [70] utilizes discrete encoding to learn the prior and employs vector quantization methods to prevent the latents from becoming uninformative.

*3) Diffusion Probabilistic Modeling:* Compared to GANs and VAEs, a new branch of generative models, diffusion models [27], [81], [82] have become dominant in many tasks such as text-to-image generation or text-to-video generation. The core idea of diffusion modeling is to learn the transformation between the real data distribution $q(x_0)$ and a standard Gaussian distribution $q(x_T)$.

We briefly introduce the denoising diffusion probabilistic model (DDPM), which includes the forward and backward processes. In the forward process, given a real data sample $x_0$, it will go through a Markov process with more and more random Gaussian noise added to the sample as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), t = 0, 1, \cdots, T \quad (4)$$

where $t$ is the time step, $T$ is usually large so that $x_T$ is close to a Gaussian noise, and $\beta_t$ is a parameter to control the noise schedule. Conversely, to achieve generation from random noise, what DDPM does in the backward process is to learn the following distribution:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (5)$$

where a neural network parameterized by $\theta$ is designed to predict the less noisy image $x_{t-1}$. Then, with this denoising network $\theta$, we can denoise from a random noise $x_T$ step by step until we get a clean data sample $x_0$, which could be an image or a video, etc.

**Remark.** GANs, VAEs, and diffusion models are all generative models. Compared to GANs, which train both the generator and discriminator, the diffusion models have explicit probabilistic modeling and only train a denoising network $\epsilon_\theta$, which is more stable. Similarly, VAEs train both an encoder and a decoder. Moreover, diffusions denoise for each image $T$ times in the training phase, resulting in $T$ variants of each image as augmentation. These augmented images in turn help the denoising network to better model the data distribution $p_\theta(x_0)$, leading to better generation results.

*4) Latent Diffusion Model:* As shown in Eq. (4) and Eq. (5), the denoising process of diffusion models is conducted on the pixels of each image in an iterative manner, which results in high computational cost, especially when the generated image is high-resolution. To tackle this problem, the latent diffusion model (LDM) [27] proposed to conduct the diffusion process in the latent space instead of the pixel space. The framework comparison between the pixel-level diffusion model and LDM is shown in Fig. 5. To reduce the computational cost, LDM utilizes the encoder of VQGAN [25] to compress the image into the latent space, $z = E(x)$, which has a much lower dimension than the original image. Then, the diffusion process in Eq. (4) and Eq. (5) will be conducted in the latent space.

Note that there is an additional input $c$ of the denoising network that is for conditional generation, e.g., as for the text-to-image generation task, $c$ could be the representation of the text prompt [83]. Also, $c$ could be other conditions, such as layout [84] or semantic maps [85]. Since most computation, including the training and iterative inference, is conducted in the lower-dimension latent space, the LDM model exhibits high efficiency. Therefore, most text-to-image and text-to-video models adopt the LDM structure.

*5) Flow Matching:* Compared with diffusion models such as DDPM, Flow Matching [86] represents a new paradigm in generative modeling, built upon Continuous Normalizing Flows (CNFs). It introduces a simple yet intuitive training objective that learns to approximate a target vector field, which defines a probability path transforming noise samples into data samples. In this way, diffusion processes can be viewed as special cases within the broader Flow Matching framework.

Let $x_1$ denote a random variable drawn from an unknown data distribution $q(x_1)$. We define a probability path $p_t$ such that $p_0 = p$ is a simple distribution, e.g., the standard normal distribution $p(x) = \mathcal{N}(x|0, I)$, and $p_1$ approximates the data distribution $q$. The goal of Flow Matching is to learn a vector field that aligns the model's probability path with this target path from $p_0$ to $p_1$.

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,p_t(x)}\|v_t(x) - u_t(x)\|^2, \tag{6}$$

where $p_t(x)$ denotes the target probability density path, $u_t(x)$ is the corresponding vector field, and $v_t(x, \theta)$ is the learnable CNF vector field parameterized by $\theta$. Here $t \sim \mathcal{U}[0, 1]$ is the uniform distribution, and $x \sim p_t(x)$. In essence, the Flow Matching loss trains the neural vector field $v_t$ to regress toward the target field $u_t$. When the loss approaches zero, the learned CNF model successfully reproduces the probability path $p_t(x)$.

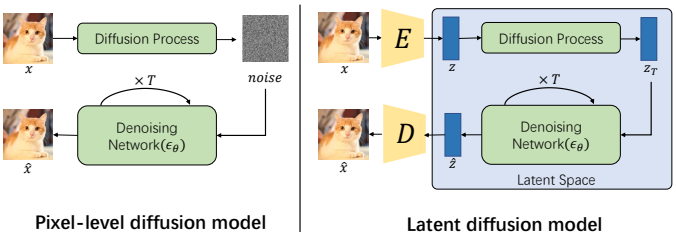


Fig. 5. Comparison between pixel-level and latent diffusion models.

## B. Text-to-Image Generation

As mentioned in the preliminary part, diffusion models can be broadly categorized into two branches: pixel-based and latent-based [87]. In the early development stage, the denoising process is typically applied directly in the pixel space. For instance, GLIDE [88] is a pioneering work in photorealistic image generation with text guidance, using a 3.5 billion parameter diffusion model that employs a text encoder to condition on natural language descriptions. GLIDE also explores the use of CLIP guidance and classifier-free guidance in diffusion models, finding that classifier-free guidance produces higher-quality images. Besides, Imagen [89] follows GLIDE and adopts classifier-free guidance for its pixel-based diffusion model. The key difference between them is that GLIDE trains a text encoder and a diffusion model together with text-image pairs, while Imagen utilizes pretrained and frozen large transformer language models, leveraging their strong text understanding capabilities to enhance sample fidelity and image-text alignment.

However, directly operating in pixel space requires substantial computational resources, which leads to the appearance of latent-based diffusion models. A milestone in this area is Stable Diffusion [90], which introduces the concept of latent diffusion model to strike a near-optimal balance between complexity reduction and detail preservation. It incorporates a pretrained VQGAN to compress images from pixel space into semantic latent space. Compared to pixel-based diffusion methods, Stable Diffusion not only achieves competitive performance across multiple image generation tasks but also significantly reduces both training and inference costs. Another notable example of a latent-based model is DALL-E2 [91], which combines a CLIP model and a diffusion model to enable zero-shot text-guided image generation. DALL-E2 consists of a CLIP image encoder and a diffusion decoder that inverts the encoder, allowing for explicit generation of image representations. This approach improves image diversity while maintaining photorealism and caption similarity.

GLIDE [88], Imagen [89], Stable Diffusion [90], and DALL-E2 [91] are all pioneering works that represent different technological pathways in the field of text-to-image generation. These models have greatly inspired subsequent research and development [92]–[94]. Despite their differences, some common trends have emerged in their development. First, latent-based diffusion methods have become increasingly prevalent due to their advantages in conserving computational resources and generating high-quality images. Second, compared to classifier guidance [95], classifier-free guidance [96]

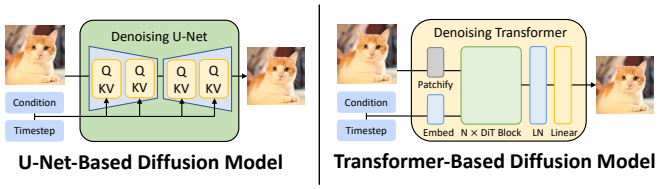


Fig. 6. Comparison between U-Net-based diffusion model and Transformer-based diffusion model.

is widely adopted in these works, where the label in a class-conditional diffusion model is replaced with a null label at a fixed probability during training. Third, U-Net traditionally serves as the backbone of the diffusion model, facilitating denoising and the gradual generation of high-quality images.

Despite its advantages in high-resolution image generation, U-Net's specific structures, such as ResBlocks and convolutional operations, limit its scalability. In contrast, Transformers, which are better suited to handle larger-scale data and tasks, are emerging as strong contenders to U-Net. The Diffusion Transformer (DiT) [97] represents a class of diffusion models that replaces the commonly used U-Net backbone with a transformer backbone, as shown in Fig. 6. This approach is supported by empirical findings suggesting that the U-Net inductive bias is not crucial to the performance of diffusion models. Additionally, utilizing a transformer backbone enables the diffusion model to leverage the best practices of transformers, such as architectural design and training paradigms, along with their good properties, such as scalability, robustness, and efficiency. Specifically, DiT adheres to the foundation of the Latent Diffusion Model (LDM) framework and emulates the design of the Vision Transformer (ViT) by introducing a comprehensive DiT design space, including patch size, transformer block architecture, and model size. The first layer of DiT, termed patchify, converts the spatial input into a sequence of tokens by linearly embedding each patch. Following the patchify step, the input tokens are processed through a sequence of transformer blocks that incorporate conditioning, such as time and label. The proposed transformer design includes adaptive layer norm (adaLN) block, cross-attention block, and in-context conditioning block. After the final block, a transformer decoder translates the image tokens into output predictions. The difference between U-Net-based and Transformer-based diffusion models is illustrated in Fig. 6.

The three distinct transformer blocks are the core modules of DiT, representing different ways to interact with multi-modal information, including images, timestep, and conditions. Their designs are inspired by the standard ViT block design but incorporate small yet significant modifications. As illustrated in Fig. 7, these blocks differ in how the image latent interacts with the conditioning information. The adaLN block follows the adaptive normalization layers in GANs, replacing the standard normalization layers in transformer blocks. The scale and shift parameters in this block are determined by the sum of the embedding vectors of timestep and condition. This block adds the least Gflops to the model. The cross-attention block introduces an additional multi-head cross-attention layer, serving as the interaction module between the image latent and

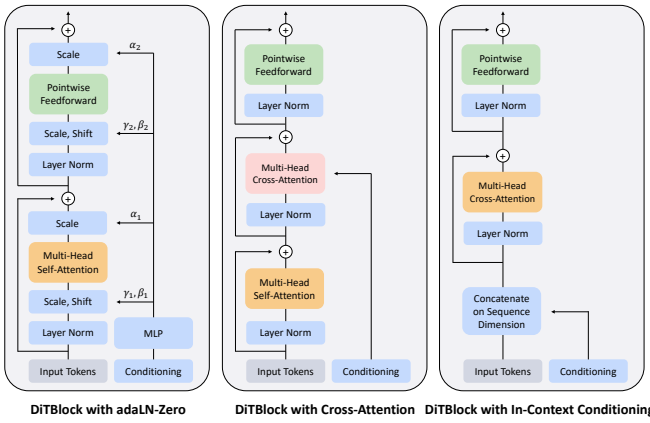


Fig. 7. Comparison between different DiT blocks from [97].

the timestep and condition. This block adds the most Gflops to the model. The in-context conditioning block treats the tokens from the timestep and condition in the same way as image tokens, concatenating them along the sequence dimension. This block introduces a moderate amount of Gflops.

Following the development of DiT [97], a growing number of works are exploring variants of diffusion transformers with improved performance. For instance, CrossDiT [98] combines the adaLN-zero DiT block and cross-attention DiT block. It simplifies adaLN-zero layers to adaLN-single layers by removing label conditioning and using only time conditioning for scale and shift control. It incorporates text embeddings from T5 [99] into the multi-head cross-attention layer. Another notable variant is MM-DiT [100], which integrates the adaLN-zero DiT block and in-context conditioning DiT block. This model uses text embeddings from CLIP and timestamps to condition the network, employs two separate sets of weights for image and condition modalities, and concatenates image and condition for the attention operation. Empirical experiments show that both CrossDiT and MM-DiT outperform the vanilla DiT in terms of validation loss, CLIP score, and FID.

The designs of diffusion transformer variants are distinct from each other, but they basically derive from the three core architectures proposed by DiT: the adaLN-zero block, the cross-attention block, and the in-context conditioning block. Currently, MM-DiT, which combines the adaLN-zero block with in-context conditioning, represents the state-of-the-art architecture. Its advantage lies in training the text modality iteratively alongside the diffusion process in an in-context manner rather than keeping it frozen, which produces a more diverse semantic space.

### C. Text-to-Video Generation

Due to the success of diffusion models in text-to-image tasks, many researchers have introduced temporal information to the diffusion models and utilized the capability of generating high-quality images to conduct text-to-video models.

The most intuitive approach to utilizing the text-to-image model is modifying the self-attention mechanism, which gets the text-to-video model without any additional parameters.

Text2Video-Zero [101] is one of the pioneer works. Rather than randomly initializing the latents of all frames independently, Text2Video-Zero only samples the latent code $z_T^1$ of the first frame and applies $\Delta t$ DDIM backward steps to obtain $z_{T'}^1$. After that, Text2Video-Zero determines the global scene and a camera motion direction, proposes a warping function $W_k$ to get all $F$ frames from $z_{T'}^1$ to $z_{T'}^F$, and then performs a DDPM forward to get the initial latents. To keep the consistency among different frames, Text2Video-Zero proposes cross-frame attention, which uses keys and values from the first frame to generate the images. Latent-Shift [102] is another representative method. It proposes a novel Temporal-Shift module that splits the latents along the channel dimension and shifts the split channel along the temporal dimension to keep the consistency of all frames. These methods have fully used the powerful pretrained text-to-image models and can generate videos with much higher resolution and quality than traditional text-to-video methods using GANs and VAEs. However, rather than capturing, training, and understanding the temporal information, these methods are more like providing a class of expert knowledge that can utilize the temporal information from a human perspective. Thus, these methods enjoy high generation efficiency, but the videos generated still struggle with motion smoothness and video consistency.

To solve the problems, another kind of approaches [103]–[105] not only inherits the architecture of the T2I models but also makes efforts to introduce novel modules or modify the original structure to learn the temporal information. VDM [103] is one of the earliest works that transferred the T2I model to solve T2V tasks. VDM proposes a 3D U-Net that modifies the diffusion architecture by changing each 2D spatial convolutional layer into a 3D convolution. After that, for each spatial attention block, VDM inserts a temporal attention block that performs attention over all frames with relative position embeddings to distinguish the ordering of frames. Make-a-video [104] proposed a pseudo-3D convolutional and attention layer, which consists of a spatial 2D convolutional layer and a temporal 1D convolutional layer. Compared to 3D convolution, this approach is much more efficient while facilitating information sharing between the spatial and temporal axes. To more flexibly apply the capabilities of the T2I model, such as the customization and style transferring ability brought by LoRA, AnimateDiff [105] keeps the original architecture and only inserts a motion module after each pretrained layer. The motion module consists of an input projection layer, several temporal self-attention layers, and an output projection layer. To avoid harming the original capabilities of T2I models, AnimateDiff zero initializes the output projection layer.

As the attention-based architecture is more suitable for capturing long-range contextual relationships, some methods [106], [107] adopt a DiT-based model to generate videos. Latte [106] utilizes a video transformer as the backbone and employs a VAE to encode videos into features, which is used to extract tokens. Currently, compared to U-Net-based methods, DiT-based methods can scale to larger datasets and parameters, hence yielding relatively better performance. However, this also implies a higher consumption of computational resources. The DiT-based methods are commonly adopted in accomplishing some outstanding applications within the industry.

### D. Text-to-Speech Generation

Text-to-Speech (TTS) generation, also known as speech synthesis, is one of the most fundamental tasks in multimodal speech processing [108]. The development of TTS has evolved from a three-stage pipeline to a two-stage framework, and more recently, to end-to-end systems. Before the advent of neural networks, TTS systems typically converted text into linguistic features and then into acoustic features before decoding them into waveforms. With the introduction of neural networks, this process was simplified, where text only needs to be transformed into either linguistic or acoustic representations. Most recent diffusion-based TTS models adopt a two-stage approach: an acoustic model first generates acoustic features, which are then converted into waveforms using a vocoder. Moreover, several studies explore end-to-end TTS frameworks that directly synthesize speech waveforms from text input.

For two-stage text-to-speech diffusion models, the acoustic model and vocoder are the two key components. The acoustic model converts text into acoustic representations, while the vocoder synthesizes waveforms from these features. DiffWave [109] is one of the earliest diffusion-based speech synthesis models, serving as a neural vocoder. It formulates waveform generation as a DDPM task, where a neural network learns to reverse a gradual noising process applied to real waveforms. WaveGrad [110] also functions as a vocoder, introducing a continuous-time, score-based diffusion approach that models a gradient field to guide the denoising process, rather than relying on a discrete noise schedule. Grad-TTS [111] is a diffusion-based acoustic model that extends diffusion modeling from vocoders to full TTS systems. It generates acoustic features from text through stochastic differential equations (SDEs), enabling a non-autoregressive acoustic modeling framework. Diff-TTS [112] is another diffusion-based acoustic model that further advances speech synthesis by formulating the entire acoustic modeling process as a deterministic or stochastic denoising procedure.

Compared with two-stage approaches, end-to-end text-to-speech diffusion models reduce error propagation and produce higher-quality speech, becoming the mainstream development direction. For example, WaveGrad 2 [113] discards the two-stage design of WaveGrad [110] and adopts an end-to-end framework that directly generates audio from a phoneme sequence. Moreover, recent systems such as TTS-1 [114] and MiniMax-Speech [115] also follow end-to-end architectures and achieve remarkable performance in speech generation.

## IV. Unification of Understanding and Generation

Until now, we have discussed both the multi-modal LLMs and the multi-modal diffusion models, where the former works well for multi-modal understanding and the latter exhibits a powerful ability in visual generation. Then a natural question arises: could we have a unified model that can simultaneously work well for multi-modal understanding and generation? Next, we will discuss this trending problem from the following two perspectives: (i) the probabilistic modeling method, and (ii) the model architecture.

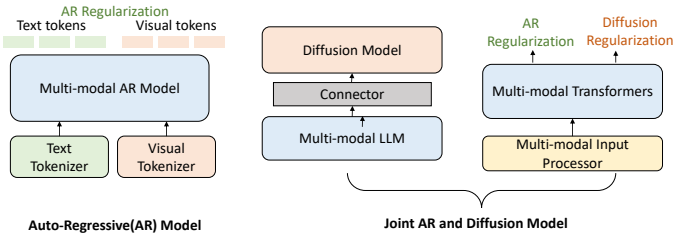


Fig. 8. Possible unified multi-modal understanding and generation frameworks with different probabilistic modeling methods.

### A. Probabilistic Modeling: Autoregressive or Diffusion?

The success of multi-modal large-language models has clearly shown the great power of autoregressive modeling for multi-modal understanding and text generation, so we believe the autoregressive method should be included. Then, the next question is how we enable the model with visual generation ability. Based on existing works in Sec. II and Sec. III, we provide the possible methods in Fig. 8, where we present the autoregressive model and the joint autoregressive and diffusion model. Next, we will elaborate on them in detail.

*1) Autoregressive (AR) Model:* Although diffusion models have become dominant in visual generation, there are still some recent attempts [3], [48], [116]–[120] on generating visual content in an autoregressive manner. These works will first try to map the input images and text into discrete tokens, respectively. Particularly, the images are discretized with visual tokenizers such as VQGAN or VQ-VAE. Then the mixed text and visual tokens will be fed into a multi-modal autoregressive model. After that, the model will output the mixed text and visual tokens. Also, some special tokens such as $< soi >, < eoi >$ are used to indicate the start of the image tokens and the end of the image tokens. Then the generated text tokens will deliver how the model understands the input multi-modal information, and the visual tokens will be sent to the decoder of the VQ-VAE or VQGAN to reconstruct images. Therefore, the autoregressive model can be used for both understanding and visual generation.

**Remark.** Despite these efforts, the autoregressive method is far from perfect — it basically assumes the existence of a causal structure and causal attention, where previous tokens are used to predict next tokens. However, this is not suitable for image generation because it is difficult to determine, which visual token should be the first and which one should be the last. Therefore, a recent work VAR [121] tries to use the next-scale prediction paradigm to generate images, where the lower-resolution images are regarded as previous tokens to predict (next) higher-resolution images. Unfortunately, the scaling ability is still not verified in multi-modal understanding and generation, and the model achieves a 1.73 FID score on the ImageNet [122] benchmark for generation, falling behind the diffusion model [123] which has a 1.35 FID score. In general, joint AR and diffusion models outperform unified AR models on visual generation tasks. For instance, EMU3 [48] and Janus-Pro [124], both unified AR models, achieve 0.66 and 0.80 on the GenEval benchmark, respectively. In contrast, joint AR-diffusion models such as Mogao [125] and Bagel [126] reach 0.89 and 0.88, demonstrating the advantages of combining AR and diffusion components for visual generation.

*2) Joint Autoregressive and Diffusion Model:* Considering the impressive visual generation ability of the diffusion model, a more natural way for unified multi-modal understanding and generation is to combine the autoregressive and diffusion models. In Fig. 8, we present two kinds of possible frameworks.

The first one is that we have a pretrained diffusion model for visual generation and a multi-modal LLM for multi-modal understanding. We then connect these two components, forming what we call Connector-based Joint Models. Regarding how to connect these two parts, many existing works [127]–[129] directly use the LLM as the controller and the diffusion model as a tool for visual generation, which is a common paradigm in tool learning. Although works like tool learning can enable the models with visual generation abilities, they easily suffer from generation failure when meeting multi-modal generation conditions. For example, when we want to generate "a specific girl (described with a given image) and a specific dog (described with a given image) playing on the grass", the tools available are only SOTA text-to-image models. They will fail to guarantee that the specific girl and dog occur in the generated image. In fact, there are many conditions that cannot be described with only text, and this kind of tool-learning method will fail. To tackle the problem, a more advanced way is to train a learnable connector [130]–[133], which aligns the diffusion model and the multi-modal LLM in the same space, similar to the training paradigm of the alignment module in multi-modal LLM. The alignment process enables the diffusion model to receive the LLM output multi-modal embeddings as conditions instead of pure text descriptions, thus achieving multi-modal generation. However, this paradigm inherits the limitations of alignment architecture. The multi-modal LLM and the diffusion model are pretrained respectively. The performance of the unified model will be limited by each model. Additionally, from an intuitive perspective, multi-modal understanding and multi-modal generation should not be independent tasks but rather two related tasks that could share knowledge. To train such a model, both the MLLM and the diffusion model can be frozen, and only the connector is trained. This maximally preserves the capabilities of the two models, but the information bottleneck between them can be particularly severe. Alternatively, one or both of the models can be included in training, but this requires a larger amount of data and computational resources to ensure that the original abilities of the models are not compromised. For example, in Qwen-Image [134], the MLLM is kept frozen while the diffusion model is trained on a large dataset. This preserves the full capability of the MLLM while endowing it with strong generative ability.

The second possible model is a unified multi-modal-transformer framework as shown in Fig. 8, where we do not rely on two pretrained models, but try to use a single model trained with both diffusion and autoregressive regularizations, which we refer to as Autoregressive-Diffusion Joint Models. The multi-modal input processor will first transform the multi-modal data into sequences that can be received by the transformers. Then the multi-modal transformer will try to

learn the multi-modal knowledge for both understanding and generation. Specifically, the training objectives are designed differently for each modality: text prediction uses an autoregressive regularization (computed token-wise), while image prediction uses a diffusion regularization (computed over the entire image, covering multiple patches). During inference, the model dynamically switches between language modeling and diffusion modes. In language modeling mode, it samples tokens sequentially; upon generating the BOI token, it switches to diffusion mode, appending a sequence of pure noise patches corresponding to the target image size, and gradually generates the image through T-step denoising iterations. At each step, the model predicts the noise based on the current image representation and updates the patch sequence until denoising is complete. The EOI token is then appended, and the model switches back to language modeling mode. Note that this is a transformer-like model but not necessarily an LLM. This is because when using transformers to generate visual content, the full-attention mechanism is usually adopted. In contrast, the attention mechanism adopted by LLM is causal and uni-directional. Therefore, an adaptive or mixed attention mechanism might be designed. This perspective is verified in TransFusion [135] and Show-o [136]. The difference between Transfusion and Show-o mainly lies in the diffusion model, where TransFusion adopts continuous diffusion that is similar to current visual diffusion models, but Show-o adopts masked generative modeling [137], which could be regarded as discrete diffusion regularization. Therefore, Show-o still relies on a pixel-level visual tokenizer for image generation but might trade off some understanding ability. Additionally, these two works are primary attempts at combining autoregressive and diffusion modeling methods in a single transformer-like model. There still exist several open problems regarding what the model architecture should be like, such as the multi-modal input processor or the transformer-like model, which we will discuss next.

### B. Model Architecture

Compared to previous multi-modal LLM or Diffusion models that only focus on one task, i.e., generation or understanding, the unified model itself should support multiple objectives. When it comes to understanding, the model should have the ability of conceptual abstraction and associative reasoning. In contrast, when it comes to visual generation, besides the overall concepts and their relations, pixel-level details are also important. Therefore, the unified model architecture design might be different from that of previous single-objective models. Next, we mainly discuss the possible architectures of the multi-modal input processor and the multi-modal transformers.

*1) Multi-modal input processor:* To tackle the multi-modal input text and images, two possible input processors are presented in Fig. 9. Text is consistently tackled by a text tokenizer. However, there are some differences in the visual input. In Fig. 9(a), we show the visual processor adopted by most early works, where a single visual encoder is used to process the images. Considering that the visual tokens should support the pixel-level visual generation task, early works [3], [135], [136]

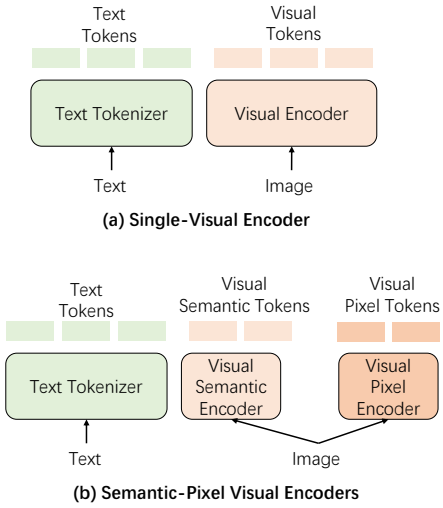


Fig. 9. Possible frameworks of the multi-modal input processor for unified multi-modal understanding and generation models.

generally adopt the single pixel-level (or patch-level) visual tokens (e.g., VQVAE). The pixel-level tokens bring challenges to the multi-modal transformer, requiring it not only to capture the relations between image patches for visual generation but also to visual abstract reasoning ability for understanding. In contrast, a possible alternative multi-modal input processor is presented in Fig. 9(b). For each image, we respectively use a semantic encoder (e.g., CLIP-ViT) and a pixel-level encoder (e.g., VQVAE) to obtain both semantic and pixel tokens. Janus [120] was the first to adopt this architecture. It introduced two separate visual encoding paths: a semantic encoder for extracting visual features in understanding tasks, and a pixel-level encoder for encoding images in generation tasks. Subsequent works, such as UniToken [138], further explored directly concatenating features of the two encoders along the sequence dimension, allowing the model to receive both types of features simultaneously for understanding and generation tasks. By using a dual-encoder approach, models can leverage both low-level pixel information and high-level semantic information, which better enhances performance on both understanding and generation tasks. Consequently, most recent works adopt this architecture. Moreover, it is a more flexible way to conduct some adaptive token selection from the semantic and pixel tokens for fine-grained understanding. We believe this would result in interesting research work.

*2) Multi-modal Transformer:* After discussing how to tackle the multi-modal input information, the next key component is the multi-modal transformer, which captures the complex relations among and within modalities. As shown in Fig. 10, on the left is a dense model, where one unified transformer is used for both multi-modal understanding and generation [45], [139]. Considering that understanding and generation might share some knowledge but their objectives are not exactly the same, it is a natural idea to utilize the mixture of experts [140] in multi-task learning as shown in (b). On the right of the figure, some of the experts share the knowledge of understanding and generation, e.g., concepts and their relations, some of the experts are good at

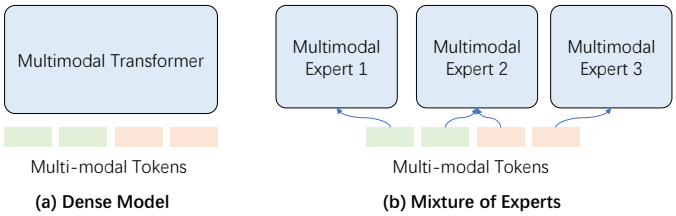


Fig. 10. Possible architectures of the multi-modal transformer.

analyzing visual details for visual generation, and other experts are good at conducting reasoning for better understanding. LlamaFusion [141] and BAGEL [126] have made preliminary explorations in this area, both using only two experts and employing hard routing. In LlamaFusion, which uses a single visual encoder, one expert is responsible for processing text tokens, while the other handles visual tokens. In contrast, BAGEL, which adopts semantic-pixel visual encoders, assigns one expert to process text tokens and visual semantic tokens, and the other to handle visual pixel tokens. Both works find that their architectures outperform dense models, indicating that unified models still face optimization challenges arising from task-specific or modality-specific learning objectives.

In Table II, we present the performance of several recent unified models. Due to large differences in model size and training data volume, a fair comparison is difficult. Regarding architecture choice: currently, there are still no large-scale Autoregressive Models trained with massive data. The latest Skywork UniPic demonstrates strong capabilities in generation and editing, but its performance on understanding tasks is not reported. In the Connector-based Joint Models category, MetaQueries, BLIP3o, and Qwen-Image all adopt Qwen2.5-vl-7B as the MLLM, resulting in similar performance on understanding tasks. However, the success of Qwen-Image indicates that increasing the scale of the diffusion model and enlarging the training dataset can significantly boost performance in generation and editing tasks. In the Autoregressive-Diffusion Joint Models category, BAGEL leverages the largest model and dataset, making it a strong competitor to Qwen-Image. Regarding the choice of visual encoder: most recent models adopt the dual encoder (Semantic-Pixel Visual Encoders) architecture, which benefits both understanding and generation tasks. Since models using MoE are still limited, it remains unclear whether MoE brings significant advantages. We hope that future work will explore this direction further.

In this section, we provide a discussion of the unified model of multi-modal generation and multi-modal understanding, from both the probabilistic modeling methods and the model architectures. Though the discussed techniques can combine with each other to form more architectures as well, there are very few attempts at the unified model design, making us believe the inspirations of many future works brought by the discussions above.

## V. Datasets

After discussing the multi-modal understanding and generation models, multi-modal text-image and text-video datasets are also important to implement multi-modal generative AI [161]. In this section, we will review the literature on the datasets for training multi-modal generative AI models. Based on the differences in data types, we divide the datasets into three categories: caption, conversation, and reasoning. In addition, many multi-modal large foundation models choose to collect the aforementioned types of data for integration and construct their own datasets. Therefore, we denote these datasets as the integration datasets.

### A. Caption Datasets

The caption dataset aims to improve basic visual and temporal description capabilities for multi-modal LLMs and provide the mapping relationship for text-to-image and text-to-video models. Commonly used text-to-image datasets include SBU Captions [162], MSCOCO [163], Conceptual Captions (CC-3M) [164], and LAION [165]. The size of these datasets ranges from 328K to 5B. Recently, MINT-1T has been proposed, comprising one trillion text tokens and three billion images [166], a 10x scale-up from existing open-source datasets, and it includes previously untapped sources such as PDFs and ArXiv papers. Text-to-video datasets include WebVid [167], InternVid [168], HD-VG-130M [169], YouCook2 [170], and TextVR [171].

The caption datasets mainly serve in the following two aspects, i.e., (i) provide knowledge for the training of generation models to generate images or videos based on the input text embedding, and (ii) use text-image datasets to align the image modality with the multi-modal LLM for understanding inputs.

### B. Conversation Datasets

The conversation dataset aims at enhancing multi-modal LLMs' capabilities for single-turn and multi-turn conversations when asking questions about the input image or video. Normally, a diverse set of questions would be asked about the visual content of the image and the video, including the object types, counting the objects, object actions, object locations, event moment, event duration, and relative positions between objects. With simple formatting reorganization, many visual QA datasets could be directly constructed as conversation datasets for multi-modal LLM training. These include basic VQA (VQAv2 [172], GQA [173]), knowledge-based VQA (OK-VQA [174], AOK-VQA [175]), OCR-based VQA (OCR-VQA [176], TextVQA [177]) and VideoQA (TGIF-QA [178], WebVidQA [179], and egocentric VQA from Ego4D [180]), which can not only improve the visual QA capabilities for multi-modal LLMs in conversations but also help the models to learn more visual and temporal knowledge.

### C. Reasoning Datasets

The above two types of datasets mainly focus on the visual content itself, normally lacking in-depth reasoning questions. Meanwhile, the reasoning datasets focus on enhancing multi-modal LLMs for diverse reasoning capacities, which normally require a step-by-step reasoning process by following rigorous logic. These include spatial reasoning (CLEVR [181]), reading comprehension (VisualMRC [182]), temporal reasoning (NExT-QA [183]), and spatiotemporal reasoning (CLEVRER [184]).

TABLE I
OVERVIEW OF MULTI-MODAL LLM, DIFFUSION, AND UNIFIED MODELS IN THIS PAPER.

| Model | Institution | Type | Classification | Publication | Year | Parameters |
|---|---|---|---|---|---|---|
| **Multi-modal LLM (MLLM)** | | | | | | |
| LLaVA [4] | Microsoft | Image LLM | Alignment | NeurIPS | 2024 | 13B |
| BLIP-2 [28] | Salesforce | Image LLM | Alignment | ICML | 2023 | 12B |
| MiniGPT-4 [34] | KAUST | Image LLM | Alignment | ICLR | 2024 | 7B |
| Qwen-VL [142] | Alibaba | Image LLM | Alignment | ArXiv | 2023 | 7B |
| Flamingo [33] | DeepMind | Image LLM | Alignment | NeurIPS | 2025 | 3B |
| Fuyu [37] | Adept | Image LLM | Early-Fusion | - | 2023 | 8B |
| Gemini [29] | Google | Image LLM | Early-Fusion | ArXiv | 2023 | - |
| Claude3 [143] | Anthropic | Image LLM | Early-Fusion | - | 2024 | - |
| VideoChat [52] | Shanghai AI Lab | Video LLM | Alignment | ArXiv | 2023 | 7B |
| VideoLLaMA [53] | Alibaba | Video LLM | Alignment | EMNLP | 2023 | 7B |
| VideoLLaMA2 [55] | Alibaba | Video LLM | Alignment | ArXiv | 2024 | 7B |
| Video-ChatGPT [54] | MBZUAI | Video LLM | Alignment | ACL | 2023 | 7B |
| LLaVA-OneVision [19] | ByteDance | Video LLM | Alignment | TMLR | 2024 | 7B |
| MiniCPM-V [144] | OpenBMB | Video LLM | Alignment | ArXiv | 2024 | 8B |
| VILA-1.5 [145] | NVIDIA | Video LLM | Alignment | ArXiv | 2023 | 7B |
| Pengi [146] | Microsoft | Speech LLM | Alignment | NeurIPS | 2023 | 1B |
| Salmonn [147] | ByteDance | Speech LLM | Alignment | ICLR | 2024 | 13B |
| Qwen-Audio [148] | Alibaba | Speech LLM | Alignment | ArXiv | 2023 | 7B |
| OSUM [149] | NPU | Speech LLM | Alignment | ArXiv | 2025 | 7B |
| VALL-E [72] | Microsoft | Speech LLM | Early-Fusion | ArXiv | 2025 | 7B |
| SpeechGPT [73] | Fudan University | Speech LLM | Early-Fusion | EMNLP | 2023 | 7B |
| AudioPaLM [74] | Google | Speech LLM | Early-Fusion | ArXiv | 2023 | 8B |
| **Diffusion Model** | | | | | | |
| GLIDE [88] | OpenAI | Text-to-Image | Pixel-Based | ICML | 2022 | 5B |
| Imagen [89] | Google | Text-to-Image | Pixel-Based | NeurIPS | 2022 | 3B |
| Stable Diffusion [90] | LMU | Text-to-Image | Latent-Based | CVPR | 2022 | 1B |
| DALL-E2 [91] | OpenAI | Text-to-Image | Latent-Based | ArXiv | 2022 | 6B |
| DiT [97] | Meta | Text-to-Image | Latent-Based | ICCV | 2023 | 1B |
| PixArt-$\alpha$ [98] | Huawei | Text-to-Image | Latent-Based | ICLR | 2025 | 1B |
| Text2Video-Zero [101] | Picsart AI | Text-to-Video | Latent-Based | ICCV | 2023 | 1B |
| Latent-Shift [102] | Meta | Text-to-Video | Latent-Based | ArXiv | 2023 | 2B |
| VDM [103] | Google | Text-to-Video | Latent-Based | NeurIPS | 2022 | - |
| Make-a-video [104] | Meta | Text-to-Video | Latent-Based | ICLR | 2024 | 10B |
| AnimateDiff [105] | Shanghai AI Lab | Text-to-Video | Latent-Based | ICLR | 2024 | 1B |
| Latte [106] | Shanghai AI Lab | Text-to-Video | Latent-Based | TMLR | 2025 | 1B |
| CogVideo [150] | Z.AI | Text-to-Video | Latent-Based | ICLR | 2023 | 15B |
| Wan [151] | Alibaba | Text-to-Video | Latent-Based | ArXiv | 2025 | 14B |
| HunyuanVideo [152] | Tencent | Text-to-Video | Latent-Based | ArVix | 2024 | 13B |
| Vidu [153] | Shengshu | Text-to-Video | Latent-Based | ArXiv | 2024 | - |
| DiffWave [109] | Baidu | Text-to-Speech | Vocoder | ICLR | 2021 | 6M |
| WaveGrad [110], [113] | Google | Text-to-Speech | Vocoder | ICLR | 2021 | 23M |
| Grad-TTS [111] | Huawei | Text-to-Speech | Acoustic Model | ICML | 2021 | 30M |
| Diff-TTS [112] | Neosapience | Text-to-Speech | Acoustic Model | Interspeech | 2021 | 13M |
| **Unified Model** | | | | | | |
| VL-GPT [116] | Tencent | Unified Model | Autoregressive | ArXiv | 2023 | 8B |
| Chameleon [3] | Meta | Unified Model | Autoregressive | ArXiv | 2024 | 7B |
| Emu2 [119] | BAAI | Unified Model | Autoregressive | CVPR | 2024 | 37B |
| Emu3 [48] | BAAI | Unified Model | Autoregressive | ArXiv | 2024 | 8B |
| LlamaGen [117] | ByteDance | Unified Model | Autoregressive | ArXiv | 2024 | 3B |
| AnyGPT [118] | Shanghai AI Lab | Unified Model | Autoregressive | ACL | 2024 | 8B |
| Janus [120] | DeepSeek | Unified Model | Autoregressive | CVPR | 2025 | 1B |
| Janus-Pro [124] | DeepSeek | Unified Model | Autoregressive | ArXiv | 2025 | 7B |
| Skywork UniPic [154] | Skywork | Unified Model | Autoregressive | ArXiv | 2025 | 2B |
| VisualGPT [127] | Microsoft | Unified Model | Joint AR-Diffusion | ArXiv | 2023 | - |
| HuggingGPT [128] | Microsoft | Unified Model | Joint AR-Diffusion | NeurIPS | 2024 | - |
| MLLM-Tool [129] | Meituan | Unified Model | Joint AR-Diffusion | WACV | 2025 | 13B |
| Kosmos-G [130] | Microsoft | Unified Model | Joint AR-Diffusion | ICLR | 2024 | 2B |
| CoDi-2 [131] | Microsoft | Unified Model | Joint AR-Diffusion | CVPR | 2024 | 8B |
| Seed-X [132] | Tencent | Unified Model | Joint AR-Diffusion | ArXiv | 2024 | 13B |
| MetaQuery [155] | Meta | Unified Model | Joint AR-Diffusion | ArXiv | 2025 | 7B |
| BLIP3o [133] | Salesforce | Unified Model | Joint AR-Diffusion | ArXiv | 2025 | 8B |
| OmniGen2 [156] | BAAI | Unified Model | Joint AR-Diffusion | ArXiv | 2025 | 7B |
| Qwen-Omni [157], [158] | Alibaba | Unified Model | Joint AR-Diffusion | ArXiv | 2025 | 30B |
| Ming-Omni [159] | Ant Group | Unified Model | Joint AR-Diffusion | ArXiv | 2025 | 7B |
| TransFusion [135] | Meta | Unified Model | Joint AR-Diffusion | ICLR | 2025 | 7B |
| Show-o [136] | NUS | Unified Model | Joint AR-Diffusion | ICLR | 2025 | 1B |
| Show-o2 [160] | NUS | Unified Model | Joint AR-Diffusion | ArXiv | 2025 | 7B |
| LlamaFusion [141] | Meta | Unified Model | Joint AR-Diffusion | Arxiv | 2024 | 8B |
| Mogao [125] | ByteDance | Unified Model | Joint AR-Diffusion | Arxiv | 2025 | 7B |
| BAGEL [126] | ByteDance | Unified Model | Joint AR-Diffusion | Arxiv | 2025 | 7B |

TABLE II
COMPARISON OF RECENT MULTI-MODAL MODELS ACROSS UNDERSTANDING, GENERATION, AND EDITING BENCHMARKS.

| Model | Date | Params | Data | Dual Encoder | MoE | Understanding | | | Generation | | | Editing | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | MMBench | MMMU | MM-Vet | WISE | GenEval | DPGBench | ImgEdit | GEdit-Bench-EN |
| GPT-4o | 2025.3 | - | - | - | - | 86.0 | 70.7 | - | 0.80 | 0.89 | 86.23 | 4.20 | 7.53 |
| **Autoregressive Models** | | | | | | | | | | | | | |
| Emu3 [48] | 2024.9 | 8B | - | × | × | 58.5 | 31.6 | 37.2 | 0.39 | 0.66 | 80.6 | - | - |
| Janus-Pro [124] | 2025.1 | 7B | 144M | ✓ | × | 79.2 | 41.0 | 50.0 | 0.35 | 0.80 | 84.19 | - | - |
| Skywork UniPic [154] | 2025.8 | 2B | 130M | ✓ | × | - | - | - | - | 0.86 | 85.50 | 3.49 | 5.83 |
| **Connector-based Joint Models** | | | | | | | | | | | | | |
| MetaQueries [155] | 2025.4 | 7B+1.6B | 25M | ✓ | × | 83.5 | 58.6 | 66.6 | 0.55 | 0.80 | 82.05 | - | - |
| BLIP3o [133] | 2025.5 | 7B+1.4B | 25M | ✓ | × | 83.5 | 50.6 | 66.6 | 0.62 | 0.84 | 81.6 | - | - |
| OmniGen2 [156] | 2025.6 | 3B+4B | 66M | ✓ | × | 79.1 | 53.1 | 61.8 | - | 0.80 | 83.57 | 3.44 | 6.42 |
| Qwen-Image [134] | 2025.8 | 7B+20B | >1000M | ✓ | × | 83.5 | 58.6 | 67.1 | - | 0.87 | 88.32 | 4.27 | 7.56 |
| **Autoregressive-Diffusion Joint Models** | | | | | | | | | | | | | |
| Mogao [125] | 2025.5 | 7B | - | ✓ | ✓ | 75.0 | 44.2 | - | - | 0.89 | 84.33 | - | - |
| BAGEL [126] | 2025.5 | 14B | 1600M | ✓ | ✓ | 85.0 | 55.3 | 67.2 | 0.52 | 0.88 | 85.07 | 3.20 | 6.52 |
| Show-o2 [160] | 2025.6 | 7B | 66M | ✓ | × | 79.3 | 48.9 | - | - | 0.76 | 86.14 | - | - |

TABLE III
COMMON DATASETS

| Dataset type | Modalities | Datasets |
|---|---|---|
| Captions | Text-Image | SBU Captions [162], MSCOCO [163], CC-3M [164], LAION [165], MINT-1T [166] |
| | Text-Video | WebVid [167], InternVid [168], HD-VG-130M [169], YouCook2 [170], TextVR [171] |
| Conversation | Text-Image | VQAv2 [172], GQA [173], OK-VQA [174], AOK-VQA [175], OCR-VQA [176], TextVQA [177] |
| | Text-Video | TGIF-QA [178], WebVidQA [179], EgoQA [180] |
| Reasoning | Text-Image | CLEVR [181], VisualMRC [182] |
| | Text-Video | NExT-QA [183], CLEVRER [184] |
| Intergration | Text-Image | LLaVA-Instruct [32] |
| | Text-Video&Image | Video-LLaVA [139], VideoChat2 [185], VideoLLaMa2 [55] |

### D. Integration Datasets

Due to the strong generalization ability of multi-modal LLMs, their training data is not limited to only one single task, such as caption, conversation, or reasoning, instead requiring comprehensive pretraining for both simple and complex visual modal tasks. Therefore, many multi-modal large model works often do not use a single visual task dataset. Instead, they select subsets of several datasets from each category mentioned above for integration and adjustment, forming instruction training data that employs both image and video data for different visual modal tasks. For visual instruction tuning, LLaVA [32] is the first multi-modal LLM, which i) leverages text-only GPT-4 [186] to expand the existing bounding box, and ii) employs caption dataset (e.g., MSCOCO [163]) as multi-modal instruction tuning data. In addition, Liu et al. propose LLaVA-Instruct, which is built on a subset of the CC-3M dataset and contains 58k in conversations, 23k in detailed descriptions, as well as 77k in complex reasoning records. Following the development of visual instruction tuning, many video LLMs such as Video-LLaVA [139], VideoChat2 [185], and VideoLLaMa2 [55], are proposed, utilizing the combination of caption, conversation, and reasoning datasets under both text-image and text-video modalities.

## VI. FUTURE DIRECTIONS

Last but not least, we explore challenging problems deserving further investigation and share our insights on promising future directions for multi-modal generative AI.

### A. Unified Model for Video Understanding and Generation

In Section IV, we primarily discuss the unified models for image understanding and generation. Given the large amount of video data in the wild, we believe there will be an urgent need to extend the unification to videos [187]–[189]. Among the three architectures introduced in Fig. 8, bridging the multi-modal LLM and video diffusion model with a connector [190], [191] can be achieved in a way similar to images. However, adapting the other two architectures to videos faces significant challenges due to i) the increased computational demands caused by longer sequences, as well as ii) the difficulty in learning spatiotemporal cues. For instance, in an autoregressive model, encoding individual video frames separately using a 2D visual tokenizer fails to capture the essential temporal motion information. VideoPoet [192], which employs a 3D video tokenizer [193], encodes a 17-frame video (spanning 2.125 seconds) into 1280 tokens, limiting its ability to generate longer videos. VideoLaViT [194] introduces an efficient video

representation model by decomposing videos into keyframes and temporal motions, training separate tokenizers for each of them, which significantly improves computational efficiency. However, the training cost is still too high when scaling to the large amount of web-scale video data. Similarly, using a single model trained with both diffusion and autoregressive regularizations also encounters the same challenges, where modeling complex relations such as causal attention and spatiotemporal attention within the model remains unexplored. Therefore, it deserves more effort in advancing unified generative AI for video understanding and generation.

### B. Benchmark for the Unification

On the one hand, despite some pioneering work on studying unified models [135], [136] for understanding and generation, the corresponding evaluations are conducted separately in a non-unified way. For instance, existing works use specific benchmarks for understanding tasks, such as Flickr30k [195] and VQAv2 [172], while relying on different benchmarks for generation tasks, such as MSCOCO [163] and GenEval [196]. On the other hand, a unification benchmark offers the advantage of unified metrics and rankings, providing a more comprehensive and fair assessment of model performance across both tasks. However, designing such a benchmark is challenging, as it requires a vast amount of visual data with human annotations in various forms, including labels, rankings, and natural language descriptions. More importantly, the evaluation should ideally reflect the mutual promotion between understanding and generation. In summary, the challenges for creating a unification benchmark are threefold,

1) **Dataset construction.** The visual data should be representative, diverse, and abundant, with high-quality annotations for multiple tasks.
2) **Ranking criteria.** Models should be ranked based on a combination of understanding and generation metrics, ensuring a balanced evaluation of both capabilities.
3) **Mutual promotion.** The benchmark should include datasets or tasks that effectively demonstrate how understanding and generation enhance each other.

This being the case, developing such a benchmark is crucial for pushing forward the research on the unification of understanding and generation, making it a promising area for future investigation.

### C. Multi-modal Graph Generative AI

Graph serves as a powerful and versatile data structure used to model flexible relationships and connections between entities, being capable of modeling both naturally occurring structural *instances*, e.g., protein and molecular structures, and the *relations* between entities across diverse modalities, e.g., multi-modal knowledge graphs. Therefore, we introduce the concept of **Multi-modal Graph Generative AI** as a future research direction, where 1) multi-modal information can be utilized for graph generation and 2) structural relations can be used to facilitate multi-modal content generation.

*1) Leveraging multi-modal information for graph generation:* Current multi-modal research predominantly focuses on modalities with regular structures with fixed degrees of freedom, e.g., texts (sequences) and images (grids). However, many real-world scenarios containing various modalities exhibit highly irregular structures with arbitrary degrees of freedom, e.g., protein structures [197], molecular graphs [198], scene graphs [199], etc. Accurately understanding and generating graphs across these modalities is an important direction for future research. For instance, Yao et al. [200] explore text-to-graph generation by leveraging the domain knowledge of LLMs, and Liu et al. [201] explore text-to-molecular graph generation by integrating the graph, image, and text information. However, there are several challenges for multi-modal graph generation: i) Understanding Structures. Given the high degree of irregularity in graphs, aligning them with various modalities poses significant difficulties. ii) Generating Structures. While mainstream approaches utilize autoregressive methods for generating discrete sequence information and employ diffusion models for generating continuous grid information, the complexity of graph structures tends to necessitate new techniques for multi-modal graph generation.

*2) Leveraging structural relations to facilitate multi-modal content generation:* Traditional multi-modal learning methodologies often assume that data from different modalities are independent, whereas there can be strong intrinsic relationships across modalities in the real world [202], [203]. For example, the descriptions, chirps, and images of birds are more closely related to each other than those of other species, such as dogs and fish. Leveraging graph structure to capture these multi-modal associations may help to understand and generate new content. Ektefaie et al. [204] explore the combination of multiple data modalities via cross-modal dependencies and geometric relationships to develop multi-modal architectures, e.g., image-intensive, knowledge-grounded, and language-intensive models, in order to process diverse datasets. Yoon et al. [205] capture intricate relationships between multiple modalities through graphs to enhance pretrained language models with multi-modal context for generative tasks. Nevertheless, several challenges remain: i) The feature spaces of different modalities are heterogeneous, thus aligning them in a unified space via a multi-modal graph poses significant challenges. ii) The connections across instances from different modalities can be heterophilous, e.g., the meow of black and white cats may be very similar, but their visual appearances differ significantly, leading to varying degrees of weights regarding similarity for the connections across modalities within the multi-modal graph. iii) There may be substantial biases among different modalities, e.g., textual and visual modalities may dominate the learning process due to the ease of collecting texts and images via the Internet, while other modalities, such as acoustic perception and tactile sense, are much more difficult to collect.

Multi-modal graph generative AI holds significant potential applications: generating molecular graphs from texts can facilitate scientists in rapidly creating and editing chemical compounds with desired properties through natural language interactions, thereby accelerating the drug discovery process. Additionally, leveraging multi-modal graphs allows generative

AI systems to reference entities associated with different modalities, thereby enhancing their ability to make cross-modal associations. Therefore, we encourage efforts in promoting future research in multi-modal graph generative AI.

### D. Lightweight Multi-modal Generative AI

We define **Lightweight Multi-modal Generative AI** as the family of efficient Artificial Intelligence models capable of generating diverse types of data, including texts, images, audios, etc., while being optimized for low computational cost, fast inference, and deployment on edge devices, e.g., smartphones, IoT devices. Lightweight Multi-modal Generative AI has broad applications in various scenarios, including mobile & edge AI, IoT & embedded systems, and fast prototyping & low-cost deployment. We deem lightweight multi-modal generative AI as another promising future research direction from the following three perspectives.

**1) Lightweight diffusion models** face challenges from sampling steps, neural architectures, and tasks. The iterative sampling process is a critical limitation of diffusion models, bringing high computational cost and constraining real-time applications. Although substantial works (e.g., distillation [206], consistency model [207], [208], and flow matching [86], [209]) engage in few-steps (e.g., 4 steps) or single-step sampling, fewer-steps sampling in general may cause remarkable quality degradation. Tasks that require high quality [210], [211] still adopt multi-step sampling. Thus, it is very important to improve the few-step sampling in future investigations. Besides, the massive network architectures of diffusion models also contribute to the issue of high computational costs, which tends to be even more severe as the model size increases rapidly. Previous methods try to obtain lightweight architectures via compression techniques such as quantization [212]–[214], pruning [215], feature cache [216], [217], and neural architecture search [218], [219], etc. Although these works have achieved remarkable success, their designs are mostly tailored for the setting of multi-step sampling, either being not applicable or suffering from poor performances in few-step sampling. Therefore, exploring sampling-steps-agnostic compression methods is an important future direction as well. Moreover, traditional compression methods mainly focus on UNet-based models. Existing literature [97], [100] indicates that DiT [97] may be a better architecture, resulting in the fact that more attention will be paid to DiT-based architectures. Moreover, previous compression methods mainly focus on class-condition or text-to-image generation tasks, rarely engaging in other challenging tasks such as video generation. Exploring effective compression methods for these tasks will be meaningful as well.

**2) Lightweight multi-modal LLMs** [220], such as vision token compression [139], [221] and efficient structures (e.g., MoE [222] and Mamba [223]), have been explored in quite a few studies. However, classic powerful compression methods (e.g., quantization and pruning) are largely unexplored for multi-modal LLM. Both diffusion models [213] and LLMs [224] have gained successful compression rates via the utilization of quantization and pruning, giving us much confidence in exploring these methods for multi-modal LLMs in future research.

**3) Lightweight unified model** for multi-modal understanding and generation has been largely ignored in literature. However, given that the unified models typically have numerous parameters, there will be a huge need for the corresponding lightweight versions. As such, developing effective lightweight models for the unification of understanding and generation will be a frontier research direction with no doubt.

### E. Multi-modal Generative AI in Dynamic Environment

The multi-modal generative models discussed so far in this paper mostly do not interact with the dynamic physical world. In the future, multi-modal generative AI agents are expected to behave like humans, where they can i) perceive the multi-modal environments, ii) conduct reasoning and planning based on the perception and their current states, iii) take action to interact with the environments, and iv) improve themselves via feedbacks from the environments. A very related topic is multi-modal embodied AI [225], [226], where multi-modal LLMs are used as the controller. However, existing embodied AI methods are all parameter-fixed upon deployment, limiting their abilities to self-improve in dynamic environments, where new concepts may arise in the course of time. The new concepts may cause the Out-of-Distribution (OOD) challenges for the pretrained multi-modal generative models, which fail to take the right action under these new concepts. Therefore, future works need to deal with the problem of i) when to update the model parameters, and ii) which part of the model parameters should be updated [227], e.g., the vision or the language modules.

## VII. Conclusion

In this paper, we thoroughly discuss multi-modal generative AI, with a particular focus on multi-modal LLMs, multi-modal diffusion models, as well as the unifications of LLMs and diffusions for multi-modal understanding and generation. We comprehensively overview two well-documented multi-modal generative AI paradigms, i.e., multi-modal LLMs for multi-modal understanding and diffusion models for visual generation. We deeply analyze the underlying mathematical principles, fundamental architecture designs, and practical application scenarios, indicating how these models can contribute to different aspects of multi-modal generative AI. We further present the necessities for the unification of understanding and generation, exploring the theoretical possibilities and potential designs towards building unified models that jointly support understanding and generation. The unification may come across challenges such as trade-offs between autoregressive and diffusion modeling, as well as different choices between dense and MoE architectures. Beyond summarizing existing methods, we also highlight promising future directions and identify the corresponding key challenges. We believe that the discussions together with the insights provided in this paper will serve as a foundation for future research and foster the development of more powerful, efficient, and generalizable multi-modal generative AI.

## REFERENCES


[1] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] T. Brooks *et al.*, "Video generation models as world simulators," 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators

[3] C. Team, "Chameleon: Mixed-modal early-fusion foundation models," *arXiv preprint arXiv:2405.09818*, 2024.

[4] H. Liu *et al.*, "Visual instruction tuning," *NeurIPS*, vol. 36, 2024.

[5] S. Minaee *et al.*, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[6] W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023.

[7] Z. Liang *et al.*, "A survey of multimodel large language models," in *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 2024, pp. 405–409.

[8] J. Wu *et al.*, "Multimodal large language models: A survey," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 2247–2256.

[9] D. Caffagni *et al.*, "The revolution of multimodal large language models: a survey," *arXiv preprint arXiv:2402.12451*, 2024.

[10] F.-A. Croitoru *et al.*, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[11] L. Yang *et al.*, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[12] H. Cao *et al.*, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[13] F. Nazarieh *et al.*, "A survey of cross-modal visual content generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6814–6832, 2024.

[14] S. Li *et al.*, "Introduction to the special issue on ai-generated content for multimedia," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6809–6813, 2024.

[15] X. Zhang *et al.*, "Unified multimodal understanding and generation models: Advances, challenges, and opportunities," *arXiv preprint arXiv:2505.02567*, 2025.

[16] S. Xie *et al.*, "Towards unifying understanding and generation in the era of vision foundation models: A survey from the autoregression perspective," *arXiv preprint arXiv:2410.22217*, 2024.

[17] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[18] B. Huang *et al.*, "Vtimellm: Empower llm to grasp video moments," in *CVPR*, 2024, pp. 14 271–14 280.

[19] B. Li *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[20] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[21] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[23] A. Razavi *et al.*, "Generating diverse high-fidelity images with vq-vae-2," *NeurIPS*, vol. 32, 2019.

[24] W. Yan *et al.*, "Videogpt: Video generation using vq-vae and transformers," *arXiv preprint arXiv:2104.10157*, 2021.

[25] P. Esser *et al.*, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021, pp. 12 873–12 883.

[26] J. Yu *et al.*, "Vector-quantized image modeling with improved vqgan," *arXiv preprint arXiv:2110.04627*, 2021.

[27] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[28] J. Li *et al.*, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[29] G. Team *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[30] A. Brock *et al.*, "High-performance large-scale image recognition without normalization," in *ICML*, 2021, pp. 1059–1071.

[31] R. Girdhar *et al.*, "Imagebind: One embedding space to bind them all," in *CVPR*, 2023, pp. 15 180–15 190.

[32] H. Liu *et al.*, "Visual instruction tuning," *NeurIPS*, vol. 36, 2024.

[33] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *NeurIPS*, vol. 35, pp. 23 716–23 736, 2022.

[34] D. Zhu *et al.*, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[35] J. Cha *et al.*, "Honeybee: Locality-enhanced projector for multimodal llm," in *CVPR*, 2024, pp. 13 817–13 827.

[36] W. Li *et al.*, "Tokenpacker: Efficient visual projector for multimodal llm," *arXiv preprint arXiv:2407.02392*, 2024.

[37] A. AI, "Fuyu-8b: A unified multimodal agent for image and text understanding," https://www.adept.ai/blog/fuyu-8b, 2023.

[38] P. Jin *et al.*, "Chat-univi: Unified visual representation empowers large language models with image and video understanding," in *CVPR*, 2024, pp. 13 700–13 710.

[39] J. He *et al.*, "Multi-modal instruction tuned llms with fine-grained visual perception," in *CVPR*, 2024, pp. 13 980–13 990.

[40] T. Zhang *et al.*, "Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding," *arXiv preprint arXiv:2406.19389*, 2024.

[41] W. Wang *et al.*, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," *NeurIPS*, vol. 36, 2024.

[42] H. Fei *et al.*, "Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing," 2024.

[43] H. Liu *et al.*, "A survey on hallucination in large vision-language models," *arXiv preprint arXiv:2402.00253*, 2024.

[44] H. You *et al.*, "Ferret: Refer and ground anything anywhere at any granularity," *arXiv preprint arXiv:2310.07704*, 2023.

[45] Z. Chen *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *CVPR*, 2024, pp. 24 185–24 198.

[46] C. Jiang *et al.*, "Hallucination augmented contrastive learning for multimodal large language model," in *CVPR*, 2024, pp. 27 036–27 046.

[47] N. Stiennon *et al.*, "Learning to summarize with human feedback," *NeurIPS*, vol. 33, pp. 3008–3021, 2020.

[48] X. Wang *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[49] Y. Goyal *et al.*, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017, pp. 6904–6913.

[50] Y. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *ECCV*. Springer, 2024, pp. 216–233.

[51] Y. Tang *et al.*, "Video understanding with large language models: A survey," *arXiv preprint arXiv:2312.17432*, 2023.

[52] L. KunChang *et al.*, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[53] H. Zhang *et al.*, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023. [Online]. Available: https://arxiv.org/abs/2306.02858

[54] S. K. Muhammad Maaz, Hanoona Rasheed *et al.*, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *ArXiv 2306.05424*, 2023.

[55] Z. Cheng *et al.*, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024. [Online]. Available: https://arxiv.org/abs/2406.07476

[56] S. Bai *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[57] J. Zhu *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[58] H. Chen *et al.*, "Grounding-prompter: Prompting llm with multimodal information for temporal sentence grounding in long videos," *arXiv preprint arXiv:2312.17117*, 2023.

[59] W. Feng *et al.*, "Llm4vg: Large language models evaluation for video grounding," *arXiv preprint arXiv:2312.14206*, 2023.

[60] E. Song *et al.*, "Moviechat: From dense token to sparse memory for long video understanding," in *CVPR*, 2024, pp. 18 221–18 232.

[61] H. Liu *et al.*, "World model on million-length video and language with blockwise ringattention," *arXiv preprint arXiv:2402.08268*, 2024.

[62] P. Zhang *et al.*, "Long context transfer from language to vision," *arXiv preprint arXiv:2406.16852*, 2024.

[63] Y. Li *et al.*, "Llama-vid: An image is worth 2 tokens in large language models," *arXiv preprint arXiv:2311.17043*, 2023.

[64] Z. Wang *et al.*, "Videotree: Adaptive tree-based video representation for llm reasoning on long videos," *arXiv preprint arXiv:2405.19209*, 2024.

[65] J. Peng *et al.*, "A survey on speech large language models for understanding," *Authorea Preprints*, 2025.

[66] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[67] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[68] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[69] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

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

[71] A. Défossez *et al.*, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[72] S. Chen *et al.*, "Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers," *arXiv preprint arXiv:2406.05370*, 2024.

[73] D. Zhang *et al.*, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[74] P. K. Rubenstein *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[75] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[76] J. Bao *et al.*, "Cvae-gan: fine-grained image generation through asymmetric training," in *ICCV*, 2017, pp. 2745–2754.

[77] C. Vondrick *et al.*, "Generating videos with scene dynamics," *NeurIPS*, vol. 29, 2016.

[78] D. P. Kingma *et al.*, "Auto-encoding variational bayes," in *ICLR*, 2014.

[79] G. Mittal *et al.*, "Sync-draw: Automatic video generation using deep recurrent attentive architectures," in *ACM Multimedia*, 2017, pp. 1096–1104.

[80] Y. Li *et al.*, "Video generation from text," in *AAAI*, vol. 32, no. 1, 2018.

[81] J. Ho *et al.*, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[82] J. Song *et al.*, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[83] S. Reed *et al.*, "Generative adversarial text to image synthesis," in *ICML*, 2016, pp. 1060–1069.

[84] Y. He *et al.*, "Localized text-to-image generation for free via cross attention control," *arXiv preprint arXiv:2306.14636*, 2023.

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

[86] Y. Lipman *et al.*, "Flow matching for generative modeling," in *ICLR*, 2023.

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

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

[89] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, vol. 35, pp. 36 479–36 494, 2022.

[90] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[91] A. Ramesh *et al.*, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[92] H. Chen *et al.*, "Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation," *arXiv preprint arXiv:2305.03374*, 2023.

[93] ——, "Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning," *arXiv preprint arXiv:2311.00990*, 2023.

[94] ——, "Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control," *arXiv preprint arXiv:2405.12796*, 2024.

[95] P. Dhariwal *et al.*, "Diffusion models beat gans on image synthesis," *NeurIPS*, vol. 34, pp. 8780–8794, 2021.

[96] J. Ho *et al.*, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[97] W. Peebles *et al.*, "Scalable diffusion models with transformers," in *ICCV*, 2023, pp. 4195–4205.

[98] J. Chen *et al.*, "Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *arXiv preprint arXiv:2310.00426*, 2023.

[99] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[100] P. Esser *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first ICML*, 2024.

[101] L. Khachatryan *et al.*, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *ICCV*, 2023, pp. 15 954–15 964.

[102] J. An *et al.*, "Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation," *arXiv preprint arXiv:2304.08477*, 2023.

[103] J. Ho *et al.*, "Video diffusion models," *NeurIPS*, vol. 35, pp. 8633–8646, 2022.

[104] U. Singer *et al.*, "Make-a-video: Text-to-video generation without text-video data," in *ICLR*, 2024.

[105] Y. Guo *et al.*, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," in *ICLR*, 2024.

[106] X. Ma *et al.*, "Latte: Latent diffusion transformer for video generation," *arXiv preprint arXiv:2401.03048*, 2024.

[107] S. Chen *et al.*, "Gentron: Diffusion transformers for image and video generation," in *CVPR*, 2024, pp. 6441–6451.

[108] C. Zhang *et al.*, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai," *arXiv preprint arXiv:2303.13336*, 2023.

[109] Z. Kong *et al.*, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[110] N. Chen *et al.*, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[111] V. Popov *et al.*, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International conference on machine learning*. PMLR, 2021, pp. 8599–8608.

[112] M. Jeong *et al.*, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409*, 2021.

[113] N. Chen *et al.*, "Wavegrad 2: Iterative refinement for text-to-speech synthesis," *arXiv preprint arXiv:2106.09660*, 2021.

[114] O. Atamanenko *et al.*, "Tts-1 technical report," *arXiv preprint arXiv:2507.21138*, 2025.

[115] B. Zhang *et al.*, "Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder," *arXiv preprint arXiv:2505.07916*, 2025.

[116] J. Zhu *et al.*, "Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation," *arXiv preprint arXiv:2312.09251*, 2023.

[117] P. Sun *et al.*, "Autoregressive model beats diffusion: Llama for scalable image generation," *arXiv preprint arXiv:2406.06525*, 2024.

[118] J. Zhan *et al.*, "Anygpt: Unified multimodal llm with discrete sequence modeling," *arXiv preprint arXiv:2402.12226*, 2024.

[119] Q. Sun *et al.*, "Generative multimodal models are in-context learners," in *CVPR*, 2024, pp. 14 398–14 409.

[120] C. Wu *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," in *CVPR*, 2025, pp. 12 966–12 977.

[121] K. Tian *et al.*, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *arXiv preprint arXiv:2404.02905*, 2024.

[122] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.

[123] J. Yao *et al.*, "Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models," *arXiv preprint arXiv:2501.01423*, 2025.

[124] X. Chen *et al.*, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025.

[125] C. Liao *et al.*, "Mogao: An omni foundation model for interleaved multi-modal generation," *arXiv preprint arXiv:2505.05472*, 2025.

[126] C. Deng *et al.*, "Emerging properties in unified multimodal pretraining," *arXiv preprint arXiv:2505.14683*, 2025.

[127] C. Wu *et al.*, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023.

[128] Y. Shen *et al.*, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *NeurIPS*, vol. 36, 2024.

[129] C. Wang *et al.*, "Tool-lmm: A large multi-modal model for tool agent learning," *arXiv preprint arXiv:2401.10727*, 2024.

[130] X. Pan *et al.*, "Kosmos-g: Generating images in context with multimodal large language models," *arXiv preprint arXiv:2310.02992*, 2023.

[131] Z. Tang *et al.*, "Codi-2: In-context interleaved and interactive any-to-any generation," in *CVPR*, 2024, pp. 27 425–27 434.

[132] Y. Ge *et al.*, "Seed-x: Multimodal models with unified multi-granularity comprehension and generation," *arXiv preprint arXiv:2404.14396*, 2024.

[133] J. Chen *et al.*, "Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset," *arXiv preprint arXiv:2505.09568*, 2025.

[134] C. Wu *et al.*, "Qwen-image technical report," *arXiv preprint arXiv:2508.02324*, 2025.

[135] C. Zhou *et al.*, "Transfusion: Predict the next token and diffuse images with one multi-modal model," *arXiv preprint arXiv:2408.11039*, 2024.

[136] J. Xie *et al.*, "Show-o: One single transformer to unify multimodal understanding and generation," *arXiv preprint arXiv:2408.12528*, 2024.

[137] L. Yu *et al.*, "Magvit: Masked generative video transformer," in *CVPR*, 2023, pp. 10 459–10 469.

[138] Y. Jiao *et al.*, "Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3600–3610.

[139] B. Lin *et al.*, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[140] R. A. Jacobs *et al.*, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[141] W. Shi *et al.*, "Llamafusion: Adapting pretrained language models for multimodal generation," *arXiv preprint arXiv:2412.15188*, 2024.

[142] J. Bai *et al.*, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[143] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," https://claude.ai/, 2024.

[144] Y. Yao *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[145] J. Lin *et al.*, "Vila: On pre-training for visual language models," 2023.

[146] S. Deshmukh *et al.*, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[147] C. Tang *et al.*, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[148] Y. Chu *et al.*, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[149] X. Geng *et al.*, "Osum: Advancing open speech understanding models with limited resources in academia," *arXiv preprint arXiv:2501.13306*, 2025.

[150] W. Hong *et al.*, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," in *ICLR*, 2023.

[151] T. Wan *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[152] W. Kong *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[153] F. Bao *et al.*, "Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models," *arXiv preprint arXiv:2405.04233*, 2024.

[154] P. Wang *et al.*, "Skywork unipic: Unified autoregressive modeling for visual understanding and generation," *arXiv preprint arXiv:2508.03320*, 2025.

[155] X. Pan *et al.*, "Transfer between modalities with metaqueries," *arXiv preprint arXiv:2504.06256*, 2025.

[156] C. Wu *et al.*, "Omnigen2: Exploration to advanced multimodal generation," *arXiv preprint arXiv:2506.18871*, 2025.

[157] J. Xu *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[158] ——, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[159] I. AI *et al.*, "Ming-omni: A unified multimodal model for perception and generation," *arXiv preprint arXiv:2506.09344*, 2025.

[160] J. Xie *et al.*, "Show-o2: Improved native unified multimodal models," *arXiv preprint arXiv:2506.15564*, 2025.

[161] W. Zhu *et al.*, "Multimedia big data computing," *IEEE multimedia*, vol. 22, no. 3, pp. 96–c3, 2015.

[162] V. Ordonez *et al.*, "Im2text: Describing images using 1 million captioned photographs," *NeurIPS*, vol. 24, 2011.

[163] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

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

[165] C. Schuhmann *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NeurIPS*, vol. 35, pp. 25 278–25 294, 2022.

[166] A. Awadalla *et al.*, "Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens," *arXiv preprint arXiv:2406.11271*, 2024.

[167] M. Bain *et al.*, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *ICCV*, 2021, pp. 1728–1738.

[168] Y. Wang *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," in *The Twelfth ICLR*.

[169] W. Wang *et al.*, "Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation," 2023.

[170] L. Zhou *et al.*, "Towards automatic learning of procedures from web instructional videos," in *AAAI*, vol. 32, no. 1, 2018.

[171] W. Wu *et al.*, "A large cross-modal video retrieval dataset with reading comprehension," *Pattern Recognition*, vol. 157, p. 110818, 2025.

[172] Y. Goyal *et al.*, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017, pp. 6904–6913.

[173] D. A. Hudson *et al.*, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019, pp. 6700–6709.

[174] K. Marino *et al.*, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *CVPR*, 2019, pp. 3195–3204.

[175] D. Schwenk *et al.*, "A-okvqa: A benchmark for visual question answering using world knowledge," in *ECCV*. Springer, 2022, pp. 146–162.

[176] A. Mishra *et al.*, "Ocr-vqa: Visual question answering by reading text in images," in *2019 international conference on document analysis and recognition (ICDAR)*. IEEE, 2019, pp. 947–952.

[177] A. Singh *et al.*, "Towards vqa models that can read," in *CVPR*, 2019, pp. 8317–8326.

[178] Y. Jang *et al.*, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *CVPR*, 2017, pp. 2758–2766.

[179] A. Yang *et al.*, "Just ask: Learning to answer questions from millions of narrated videos," in *ICCV*, 2021, pp. 1686–1697.

[180] K. Grauman *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022, pp. 18 995–19 012.

[181] J. Johnson *et al.*, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017, pp. 2901–2910.

[182] R. Tanaka *et al.*, "Visualmrc: Machine reading comprehension on document images," in *AAAI*, vol. 35, no. 15, 2021, pp. 13 878–13 888.

[183] J. Xiao *et al.*, "Next-qa: Next phase of question-answering to explaining temporal actions," in *CVPR*, 2021, pp. 9777–9786.

[184] K. Yi *et al.*, "Clevrer: Collision events for video representation and reasoning," in *ICLR*, 2020.

[185] K. Li *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *CVPR*, 2024, pp. 22 195–22 206.

[186] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[187] H. Zhu *et al.*, "Multi-modal understanding and generation for object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[188] Z. You *et al.*, "Towards long video understanding via fine-detailed video story generation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[189] C. Jin *et al.*, "Mtartgpt: A multi-task art generation system with pre-trained transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6901–6912, 2024.

[190] S. Wu *et al.*, "Next-gpt: Any-to-any multimodal llm," *arXiv preprint arXiv:2309.05519*, 2023.

[191] H. Ye *et al.*, "X-vila: Cross-modality alignment for large language model," *arXiv preprint arXiv:2405.19335*, 2024.

[192] D. Kondratyuk *et al.*, "Videopoet: A large language model for zero-shot video generation," in *ICML*, 2024.

[193] L. Yu *et al.*, "Language model beats diffusion–tokenizer is key to visual generation," *arXiv preprint arXiv:2310.05737*, 2023.

[194] Y. Jin *et al.*, "Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization," *arXiv preprint arXiv:2402.03161*, 2024.

[195] P. Young *et al.*, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[196] D. Ghosh *et al.*, "Geneval: An object-focused framework for evaluating text-to-image alignment," *NeurIPS*, vol. 36, 2024.

[197] H.-C. Yi *et al.*, "Graph representation learning in bioinformatics: trends, methods and applications," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab340, 2022.
[198] N. Yang *et al.*, "Molecule generation for drug design: a graph learning perspective," *arXiv preprint arXiv:2202.09212*, 2022.
[199] H. Li *et al.*, "Scene graph generation: A comprehensive survey," *Neurocomputing*, vol. 566, p. 127052, 2024.
[200] Y. Yao *et al.*, "Exploring the potential of large language models in graph generation," *arXiv e-prints*, pp. arXiv–2403, 2024.
[201] P. Liu *et al.*, "Git-mol: A multi-modal large language model for molecular science with graph, image, and text," *Computers in biology and medicine*, vol. 171, p. 108073, 2024.
[202] J. Zhu *et al.*, "Multimodal graph benchmark," *arXiv preprint arXiv:2406.16321*, 2024.
[203] C. Peng *et al.*, "Learning on multimodal graphs: A survey," *arXiv preprint arXiv:2402.05322*, 2024.
[204] Y. Ektefaie *et al.*, "Multimodal learning with graphs," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 340–350, 2023.
[205] M. Yoon *et al.*, "Multimodal graph learning for generative tasks," *arXiv preprint arXiv:2310.07478*, 2023.
[206] A. Sauer *et al.*, "Adversarial diffusion distillation," *arXiv preprint arXiv:2311.17042*, 2023.
[207] Y. Song *et al.*, "Consistency models," in *ICML*, 2023, pp. 32 211–32 252.
[208] S. Luo *et al.*, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.
[209] X. Liu *et al.*, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *ICLR*, 2023.
[210] L. Tian *et al.*, "Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions," *arXiv preprint arXiv:2402.17485*, 2024.
[211] Z. Xu *et al.*, "Magicanimate: Temporally consistent human image animation using diffusion model," in *CVPR*, 2024, pp. 1481–1490.
[212] Y. Shang *et al.*, "Post-training quantization on diffusion models," in *CVPR*, 2023, pp. 1972–1981.
[213] S. Tang *et al.*, "Post-training quantization with progressive calibration and activation relaxing for text-to-image diffusion models," *arXiv preprint arXiv:2311.06322*, 2023.
[214] X. Li *et al.*, "Q-diffusion: Quantizing diffusion models," in *ICCV*, 2023, pp. 17 535–17 545.
[215] D. Zhang *et al.*, "Laptop-diff: Layer pruning and normalized distillation for compressing diffusion models," *arXiv preprint arXiv:2404.11098*, 2024.
[216] X. Ma *et al.*, "Deepcache: Accelerating diffusion models for free," in *CVPR*, 2024, pp. 15 762–15 772.
[217] P. Chen *et al.*, "Delta-dit: A training-free acceleration method tailored for diffusion transformers," *arXiv preprint arXiv:2406.01125*, 2024.
[218] S. Tang *et al.*, "Lightweight diffusion models with distillation-based block neural architecture search," *arXiv preprint arXiv:2311.04950*, 2023.
[219] L. Li *et al.*, "Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration," in *ICCV*, 2023, pp. 7105–7114.
[220] Y. Jin *et al.*, "Efficient multimodal large language models: A survey," *arXiv preprint arXiv:2405.10739*, 2024.
[221] Y. Li *et al.*, "Mini-gemini: Mining the potential of multi-modality vision language models," *arXiv preprint arXiv:2403.18814*, 2024.
[222] B. Lin *et al.*, "Moe-llava: Mixture of experts for large vision-language models," *arXiv preprint arXiv:2401.15947*, 2024.
[223] H. Zhao *et al.*, "Cobra: Extending mamba to multi-modal large language model for efficient inference," *arXiv preprint arXiv:2403.14520*, 2024.
[224] G. Xiao *et al.*, "Smoothquant: Accurate and efficient post-training quantization for large language models," in *ICML*, 2023, pp. 38 087–38 099.
[225] C. Zhang *et al.*, "Large language models for human-robot interaction: A review," *Biomimetic Intelligence and Robotics*, p. 100131, 2023.
[226] Y. Mu *et al.*, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *NeurIPS*, vol. 36, 2024.
[227] W. Zhu *et al.*, "Self-directed machine learning," *AI Open*, vol. 3, pp. 58–70, 2022.

**Xin Wang** is currently an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. He got both of his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence, machine learning and its applications. He has published over 200 high-quality research papers in top-tier conferences (ICML NeurIPS etc.) and journals (IEEE TPAMI, IEEE TIP etc.), winning three best paper awards including IEEE ICME and ACM Multimedia Asia. He is the recipient of ACM China Rising Star Award, IEEE TCMC Rising Star Award and DAMO Academy Young Fellow.

**Yuwei Zhou** is currently a Ph.D. student at the Department of Computer Science and Technology, Tsinghua University. He received his B.E. degree from the Department of Computer Science and Technology, Tsinghua University. His main research interests include machine learning, curriculum learning, and multi-multi-modal generative AI.

**Bin Huang** is currently a Ph.D. student at the Department of Computer Science and Technology, Tsinghua University. He received his B.E. degree from the Department of Computer Science and Technology, Tsinghua University. His main research interests include machine learning and multi-modal generative AI.

**Hong Chen** received B.E. from the Department of Electronic Engineering, Tsinghua University, Beijing, China in 2020. He is currently a Ph.D. candidate in the Department of Computer Science and Technology at Tsinghua University. His main research interests include machine learning, multi-modal information processing.

**Wenwu Zhu** is currently a Professor in the Department of Computer Science and Technology at Tsinghua University. He received his Ph.D. degree from New York University in 1996. His research interests are in the area of data-driven multimedia networking and Cross-media big data computing. He received eight Best Paper Awards, including ACM Multimedia 2012 and IEEE TCSVT in 2001 and 2019. He served as EiC for IEEE TMM (2017-2019) and IEEE TCSVT (2024-2025). He served in the steering committee for IEEE TMM (2015-2016) and IEEE TMC (2007-2010), respectively. He serves as General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019, respectively. He is an AAAS Fellow, ACM Fellow, IEEE Fellow, SPIE Fellow, and a member of The Academy of Europe (Academia Europaea).