Towards Unified Multimodal Large Language Models: More than A Survey

Xu Ma **Yitian Zhang** **Yun Fu**
Northeastern University, USA
ma.xu1@northeastern.edu

Abstract

The recent surge of interest in unified Multimodal Large Language Models (MLLMs) has catalyzed rapid progress toward general-purpose generation and understanding across different modalities. Despite the remarkable advancements, the field lacks a systematic and cohesive framework that connects these developments, revisits the motivations, and situates current trends within a broader landscape. In this survey, we present a comprehensive and in-depth review of unified MLLMs, offering both a methodology taxonomy and unique perspectives on the field. We begin by outlining the foundational concepts and prerequisites for understanding unified MLLMs. We then delve into designs from different aspects, including model architectures, loss functions, alignment techniques, and different representation strategies. Furthermore, we discuss persistent challenges and identify future promising directions. By bridging scattered progress and providing a consolidated view, this survey aims to foster a deeper and systematical understanding of unified MLLMs and inspire future innovations towards general multimodal intelligence.

1 Introduction

The landscape of generative modeling has evolved rapidly, with breakthroughs across different modalities like text, image, audio, and more. Autoregressive Transformer–based language models (Vaswani et al., 2017) model natural language by predicting tokens sequentially, yielding systems such as GPT (Radford et al., 2018, 2019; Brown et al., 2020; Achiam et al., 2023) and LLaMA (Touvron et al., 2023a,b; Grattafiori et al., 2024) that produce fluent, human-like text and have reshaped natural language understanding and generation. In parallel, diffusion models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022; Ramesh et al., 2022) revolutionized image synthesis by learning to reverse stochastic noise. Building on diffusion probabilistic models (Sohl-Dickstein et al., 2015; Ho et al., 2020), they now form the backbone of modern image generators. Meanwhile, technologies in other modalities like audio (Chen et al., 2024) and 3D point cloud (Wu et al., 2024c) also exhibit promising achievements. These uni-modal advances have catalyzed interest in unified multimodal large models that jointly model and generate across language, vision, and beyond, promising more general, interactive systems that reason over and produce multiple modalities in a coherent manner.

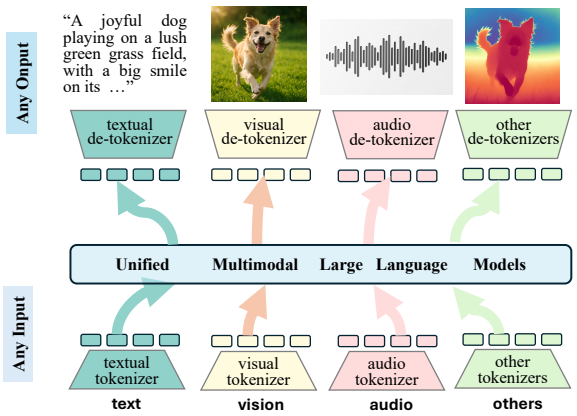

Figure 1: Unified MLLMs target any-to-any generation, which can process and generate text, vision, audio, and more modalities in a unified design.

Despite major advances in language generation via autoregressive Transformers and visual generation via diffusion models, these domains have largely progressed in isolation. Differences in input modality, architecture, and research focus have yielded a fragmented landscape, hindering models that jointly handle vision–language and other modalities generation. Still, momentum toward unified MLLMs is accelerating. Early milestones contrastive vision–language pretraining (*e.g.*, CLIP (Radford et al., 2021; Xu et al., 2023; Zhai et al., 2023)), audio-text pretraining(*e.g.*, CLAP (Wu et al., 2023; Chen et al., 2022)), and increasingly capable vision–language understanding models (Liu et al., 2023, 2024b; Li et al., 2022,

Findings of the Association for Computational Linguistics: ACL 2026, pages 37212–37230
July 2-7, 2026

2023b; Bai et al., 2023; Wang et al., 2024b; Liu et al., 2024c; Li et al., 2024a; Grattafiori et al., 2024), have begun to bridge this gap and lay the groundwork for unified MLLMs.

While these works laid the groundwork for multimodal integration, they largely generate text from multimodal inputs, falling short of truly unified multimodal generation. Building general-purpose models that both understand and generate across multiple modalities remains an open challenge. Encouragingly, recent years have seen a surge of attempts to close this gap. For example, Chameleon (Team, 2024), EMU3 (Wang et al., 2024c), Janus (Wu et al., 2025b), and OmniGen2 (Wu et al., 2025c) operate over text, images, audio, and beyond within a single architecture, spurring intense interest and rapid progress. Yet this rapid progress has produced diverse architectural strategies, training paradigms, and design philosophies. They are often without clear motivations, trade-offs, or practical implications. Additionally, the role of modality awareness in model design and the pursuit of computational efficiency are still being actively explored. These open questions sustain ambiguity around the motivations, benefits, and most promising pathways for developing unified multimodal LLMs.

In response to the rapid progress in unified multimodal modeling, this paper presents a comprehensive survey of unified MLLMs. We collect and organize key methods, foundational techniques, and related technologies that have shaped this emerging field. To illuminate the design rationale of various approaches, we introduce a structured taxonomy that categorizes existing models by architectural choices, training strategies, and modality integration paradigms. While several surveys on unified MLLMs have recently emerged (Zhang et al., 2024; Jiang et al., 2025b; Zhang et al., 2025b; Jiang et al., 2025a), the field's rapid evolution shows that earlier works often fail to capture the most recent advances. Furthermore, existing reviews are always focus on specific domains or tasks, like unifying visual generation and understanding or specific benchmarking protocols. Most importantly, a rigorous, in-depth analysis of current challenges and a strategic outlook for the field remain absent. To address these gaps, our survey moves beyond a simple taxonomy of existing methods. We trace the field's most recent development, identify critical research trends, and provide a critical analysis of current limitations and future opportunities. Our goal is to offer both a comprehensive roadmap for the community and a forward-looking perspective on the next generation of unified multimodal intelligence.

2 Background

2.1 Tokenizer for different Modalities

2.1.1 Textual Tokenizers

Tokenization is fundamental to LLMs, converting raw text into discrete tokens. To overcome the vocabulary constraints of early word-level embeddings (Mikolov et al., 2013; Pennington et al., 2014), subword algorithms like BPE (Sennrich et al., 2015), WordPiece (Wu et al., 2016), and Unigram (Kudo, 2018), became the standard for models like GPT, BERT, and T5. SentencePiece (Kudo and Richardson, 2018) later unified these into language-agnostic frameworks for multilingual scaling (Xue et al., 2020; Conneau et al., 2019). While modern LLMs (Brown et al., 2020; Achiam et al., 2023) favor byte-level encoding for Unicode robustness, "token-free" architectures like CANINE (Clark et al., 2022) and ByT5 (Xue et al., 2022) eliminate vocabulary bottlenecks entirely, albeit at higher computational costs. Most recently, the BLT (Pagnoni et al., 2025) has challenged the necessity of tokenizers altogether, though the scalability and cross-domain efficacy of such tokenizer-free approaches remain to be fully verified.

2.1.2 Visual Tokenizers

Unlike text which is human-abstracted, vision is captured as raw physical data. Visual tokenizers bridge this gap by converting images into sequences via convolutional backbones (Zhang et al., 2021) or Vision Transformers (Radford et al., 2021; Alayrac et al., 2022). These representations are either continuous, aligning features in a shared semantic space (Radford et al., 2021; Zhai et al., 2023; Bolya et al., 2025), or discrete, quantizing features into codebook indices (Van Den Oord et al., 2017; Esser et al., 2021; Ramesh et al., 2021; Sun et al., 2024a). Recent advances like lookup-free quantization (Yu et al., 2023b; Han et al., 2025) further enhance expressiveness. To mitigate the computational overhead of high-resolution, methods such as Q-Former (Li et al., 2022), token pruning, and pixel-shuffling (Ma et al., 2025b; Liu et al., 2024c) are employed to reduce token number. Effective visual tokenizer design must ultimately balance reconstruction fidelity and semantic alignment (see Table 3 in the supplementary for more details).

2.1.3 Audio Tokenizers

To integrate audio into unified MLLMs, raw waveforms or spectrograms are first mapped to language-like token sequences. Audio tokenizers perform this conversion. Recent work favors learned tokenizers, including vector-quantized models such as VQ-VAE, EnCodec (Défossez et al., 2022), and SoundStream (Zeghidour et al., 2021), which discretize audio into codebook indices for autoregressive modeling. Others derive self-supervised features from wav2vec 2.0 (Baevski et al., 2020) or HuBERT (Hsu et al., 2021) and then cluster or project them into discrete tokens. Whisper-large-v3 (Radford et al., 2023) is integrated into Qwen2.5-Omni (Xu et al., 2025). Recently, WavTokenizer (Ji et al., 2025) achieves SOTA performance with extreme compression and improved subjective quality. These tokenizers enable unified LLMs to process and generate audio for tasks like speech synthesis, audio captioning, and cross-modal understanding. Please refer a recent survey (Mousavi et al., 2025) for more details.

2.1.4 Other Tokenizers

Beyond text, visual, and audio modalities, tokenizers have also been developed for other domains such as action tokenizers (like I3D (Carreira and Zisserman, 2017) and Magma (Yang et al., 2025)), 3D point clouds (PointMLP (Ma et al., 2022), Point-MAE (Pang et al., 2023)), graph data (Graphormer (Ying et al., 2021), TokenGT (Kim et al., 2022)), molecular and protein sequences (ChemBERTa(Ahmad et al., 2022), ESM-2 (Lin et al., 2022)), and time-series sensor data (Totem (Talukder et al., 2024)). These tokenizers adapt representation strategies to the structural and statistical characteristics of their target modality, enabling LLMs to process a wider spectrum of inputs and modalities.

2.2 Large Language Models

Autoregressive (AR) modeling is the standard framework for LLMs, formulating text generation as conditional probability estimation. Given a sequence $x = (x_1, \ldots, x_T)$, the model θ is optimized by minimizing the negative log-likelihood (NLL):$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}; \theta)$, where $x_{<t}$ denotes the preceding context. This formulation is typically instantiated via causal Transformers, where an attention mask ensures that predictions for x_t depend solely on the prefix $x_{<t}$.

By leveraging teacher forcing for scalable training, this approach powers prominent models such as GPT-4 (Achiam et al., 2023) and LLaMA (Touvron et al., 2023b). Significant scaling of parameters, data, and compute within this AR framework has been shown to yield robust zero- and few-shot generalization across diverse tasks. We would suggest refer other works for a better understanding about how LLMs understand other modalities such as vision, and Sec. C and Sec. D for generation.

3 Survey on Unified MLLMs

We survey recent advances in unified MLLMs. We propose a systematic taxonomy of methods, discuss the evaluation and training of unified MLLMs.

3.1 Methodology Taxonomy

We present a detailed taxonomy of recent unified methods in Table 1. Models are ordered by release date and annotated with modalities, supported tasks, key features, and design philosophy. Overall, most work focuses on unifying text and vision, reflecting the field's current emphasis. Accordingly, we highlight vision-related choices (*e.g.*, encoders, representation strategies, and whether losses are applied to vision tokens) and focus this section on text–vision understanding and generation. Other modalities are not ignored, and we discuss these additional modalities in Sec. 3.1.6.

3.1.1 Visual Representations

Two dominant visual representation forms are considered: pixel-level (*i.e.*, low-level) and semantic-level (*i.e.*, high-level) representations. Pixel-level representations are typically extracted using VAE-based tokenizers, which emphasize reconstruction to preserve fine-grained visual details. In contrast, CLIP-based encoders are employed to capture semantic-level features that are more naturally aligned with the language embedding space. The debate between pixel-level and semantic-level representations remains an ongoing topic in the multimodal learning community. In this survey, we advocate for the adoption of semantic-level representations in unified MLLMs, guided by the Platonic Representation Hypothesis (Huh et al., 2024), which posits a shared latent reality across models, objectives, data, and modalities. Moreover, language-grounded semantic embeddings serve as an effective bridge to bind heterogeneous modalities (Zhu et al., 2023a). While we acknowledge

Model	Release Date	Modalities	Eval Tasks	100% text ability keep?	Modality Aware?	Vision Encoder	Discrete or Continuous?	Vision Loss?	Visual Decoder?
SEED (Ge et al., 2023)	07/23	Text, Image	Und / Gen	✗	✗	CLIP	Continuous	MSE loss	Diffusion
EMU (Sun et al., 2023)	07/23	Text, Image, Video	Und /Gen /Edit /Video	✗	✗	CLIP	Continuous	L2 regression	Diffusion
EMU2 (Sun et al., 2024b)	12/23	Text, Image, Video	Und /Gen /Text /Video/ Edit	✗	✗	CLIP	Continuous	L2 regression	DIffusion
SEED-X (Ge et al., 2024)	04/24	Text, Image	Und /Gen /Edit	✗	✗	CLIP	Continuous	MSE loss	Diffusion
Libra (Xu et al., 2024a)	05/24	Text, Image	Und /(Poor Gen)	✗	✗	CLIP+VQGAN	Discrete	CELoss	VQGAN
Chameleon (Team, 2024)	05/24	Text, Image	Und / Gen / Text	✗	✗	VQGAN	Discrete	CELoss	VQGAN
MoMa (Lin et al., 2024a)	07/24	Text, Image	Und / Gen / Text	✗	✓	VQGAN	Discrete	CELoss	VQGAN
Lumina-mGPT (Liu et al., 2024a)	08/24	Text, Image	Und / Gen / Edit	✗	✗	VQGAN	Discrete	CELoss	VQGAN
Transfusion (Zhou et al., 2024)	08/24	Text, Image	Und / Gen / Text	✗	✗	VAE	Continuous	Diffusion loss	VAE
Show-o (Xie et al., 2024)	08/24	Text, Image	Und / Gen / Edit	✗	✗	VQGAN	Discrete	CELoss	MaskGIT
MonoFormer (Zhao et al., 2024a)	09/24	Text, Image	Gen / Text	✗	✗	VAE	Continuous	Diffusion loss	VAE
EMU3 (Wang et al., 2024c)	09/24	Text, Image, Video	Und / Gen / Video/Edit	✗	✗	VQGAN	Discrete	CELoss	VQGAN
MIO (Wang et al., 2024d)	09/24	Text, Image, Audio	Und / Gen / Audio	✗	✗	VQGAN	Discrete	CELoss	VQGAN
Janus (Wu et al., 2025b)	10/24	Text, Image	Und / Gen	✗	✗	CLIP, VQVAE	Disc. for Gen; Con. for Und	CELoss for Gen	VQGAN
Janus-Flow (Ma et al., 2025c)	11/24	Text, Image	Und / Gen	✗	✗	CLIP, VAE	Continuous	Rectified flow	VAE
Liquid (Wu et al., 2024a)	12/24	Text, Image	Und / Gen / Text	✗	✗	VQGAN	Discrete	CELoss	VQGAN
LMFusion (Shi et al., 2024)	12/24	Text, Image	Und / Gen / Text	✓	✓	VAE	Continuous	Diffusion loss	VAE
MetaMorph (Tong et al., 2024)	12/24	Text, Image	Und / Gen	✗	✗	CLIP	Continuous	Cos sim loss	Diffusion
Janus-Pro (Chen et al., 2025c)	01/25	Text, Image	Und / Gen	✗	✗	CLIP, VQGAN	Disc. for Gen; Con. for Und	CELoss for Gen	VQGAN
UniFluid (Fan et al., 2025)	03/25	Text, Image	Und / Gen	✗	✗	CLIP, VAE	Continuous	Diffusion loss	VAE
Qwen2.5-Omni (Xu et al., 2025)	03/25	Text, Image, Audio	Und / Audio	✗	✗	CLIP	Continuous	-	-
MetaQuery (Pan et al., 2025)	04/25	Text, Image	Und / Gen	✓	✗	CLIP	Continuous	Diffusion loss	Diffusion
X-Fusion (Mo et al., 2025)	04/25	Text, Image	Und / Gen	✓	✓	VAE, CLIP	Continuous	Diffusion loss	VAE
BEGAL (Deng et al., 2025)	05/25	Text, Image	Und / Gen / Edit	✗	✓	VAE, CLIP	Continuous	Rectified Flow	VAE
OpenUni (Wu et al., 2025d)	05/25	Text, Image	Und / Gen	✓	✗	CLIP	Continuous	Diffusion loss	Diffusion
BLIP3-o (Chen et al., 2025b)	05/25	Text, Image	Und / Gen	✓	✗	CLIP	Continuous	Diffusion loss	Diffusion
UniGen (Wu et al., 2024b)	05/25	Text, Image	Und / Gen	✗	✗	VAE, CLIP	Disc. for Gen; Con. for Und	CELoss for Gen	MaskGIT
Mogao (Liao et al., 2025)	05/25	Text, Image	Und / Gen / Edit	✗	✗	CLIP+VAE	Continuous	flow matching	VAE
Ming-Omni (AI et al., 2025)	06/25	Text, Image, Audio, Video	Und / Gen	✗	✗	CLIP	Continuous	Diffusion loss	Diffusion
Ovis-U1 (Wang et al., 2025b)	07/25	Text, Image	Und / Gen / Edit	✗	✗	CLIP, VAE	Continuous	Unknown	Diffusion
Nexus-Gen (Zhang et al., 2025a)	07/25	Text, Image	Und / Gen / Edit	✗	✗	CLIP	Continuous	MSE + Cosine Loss	Diffusion
Show-o2 (Xie et al., 2025)	07/25	Text, Image, Video	Und / Gen / Video	✗	✗	VAE	Continuous	flow matching	VAE
Lumina-mGPT 2.0 (Xin et al., 2025)	07/25	Text, Image	Gen / Edit	✗	✗	VQGAN	Discrete	CELoss	VQGAN
X-Omni (Geng et al., 2025)	07/25	Text, Image	Und / Gen	✗	✗	CLIP	Discrete	CELoss	Diffusion
OmniGen2 (Wu et al., 2025c)	08/25	Text, Image	Und / Gen / Edit	✗	✗	CLIP	Continuous	Diffusion loss	VAE
UniPic (Wang et al., 2025c)	08/25	Text, Image	Gen / Edit	✗	✗	CLIP+VAE	Continuous	Diffusion loss	VAE
UniPic2-Metaquery (Wei et al., 2025)	09/25	Text, Image	Und / Gen / Edit	✓	✗	CLIP	Continuous	Diffusion loss	Diffusion
Manzano (Li et al., 2025a)	09/25	Text, Image	Und / Gen	✗	✗	CLIP	Disc. for Gen; Con. for Und	CELoss	Diffusion
UNI-X (Hao et al., 2025)	09/25	Text, Image	Und / Gen	✗	✓	VQGAN	Discrete	CELoss	VQGAN
Hunyuan3D-Omni (Hunyuan3D et al., 2025)	09/25	Image, Voxel, Point	Gen	✗	✗	VAE	Continuous	Diffusion loss	VAE
ProLLaMA (Lv et al., 2025)	09/25	Text, Protein	Gen	✗	✗	-	-	-	-
Tuna (Liu et al., 2025c)	12/25	Text, Image	Und / Gen / Edit	✗	✗	CLIP + VAE	Continuous	Diffusion loss	VAE
EMMA (He et al., 2025)	12/25	Text, Image	Und / Gen / Edit	✗	✓	CLIP + VAE	Continuous	Diffusion loss	VAE
UniHetero (Chen et al., 2025a)	12/25	Text, Image	Und	✗	✗	DINOv2	Continuous	Diffusion loss	Diffusion
JavisGPT (Liu et al., 2025a)	12/25	Text, Audio, Video	Und / Gen	✗	✗	CLIP	Continuous	Diffusion loss	Diffusion

Table 1: Overview of recent Unified MLLMs. For Eval tasks, "Und" indicates visual understanding tasks, "Gen" means visual generation tasks, "Edit" suggests visual editing tasks, "Text" means text generation evaluations. We use "/" to combine different evaluation tasks in the work. "100% text ability keep" indicate if the unified MLLMs can maintain the original backbone VLM text generation ability. For vision encoder, "CLIP, VAE" indicates parallel two vision encoders, and "CLIP+VAE" means combine the two encoders into a unified vision encoder.

the potential loss of fine visual details in CLIP-based vision tokenizers, several observations mitigate this concern. First, recent studies demonstrate that CLIP-based tokenizers retain substantial pixel-level fidelity (Sun et al., 2024b; Ge et al., 2024). Second, extensive efforts—such as VILA-U and AToken—have been made to endow CLIP encoders with reconstruction capability. Finally, complementary pixel-level refinements can be incorporated within the visual decoder, as exemplified by OmniGen2 (Wu et al., 2025c).

3.1.2 Modality-Aware Experts

Inspired by human behavior, some efforts hold the assumption that unified MLLMs should be designed as modality-aware. That is, different modalities should have different set of parameters to process, which is also termed as Mixture of Transformers (MoT) design. One of the representative work is LMFusion (Shi et al., 2024), which combines two set of Transformer parameters for text and vision separately. To bridge the knowledge from different modalities, the attention weights are calculated across all context regardless the modalities. Such modality-aware design also show strong performance in other domains, like MMDiT (Esser et al., 2024) for Diffusion text-to-image generation. This kind of design, not only consider the intrinsic differences among different modalities, but also provide strong performances. We believe such a MoT modality-aware design would exhibit promising potential in the field of unified MLLMs.

3.1.3 Attention Masking Strategies

Conventional LLMs and VLMs typically adhere to an autoregressive generation paradigm by employing a causal attention mask. However, extensive empirical studies have shown that such causal masking substantially degrades generation quality, as demonstrated in Transfusion (Zhou et al.,

2024) and VAR (Tian et al., 2024). For example, Transfusion applies causal masking to text tokens while adopting bidirectional attention for visual tokens, achieving notable improvements in vision generation performance. Further engineering and performance analyses are explored in Token-Shuffle (Ma et al., 2025b). In this survey, we posit that the bidirectional attention mechanism is inherently advantageous for visual generation—and potentially for visual understanding as well. Nevertheless, given the scalability considerations and the maturity of existing VLM pipelines, maintaining a unified autoregressive framework remains a more implementation-friendly and system-consistent choice. To mitigate the degradation in visual generation performance caused by causal masking, recent advances in multimodal positional encoding, such as MSRoPE (Wu et al., 2025a) and MRoPE (Bai et al., 2025), offer promising directions for enhancement.

3.1.4 Loss Functions

For text generation, cross-entropy loss remains the standard objective. Extending pretrained VLMs and LLMs to native image generation is still under active exploration, particularly in how to apply loss guidance to visual outputs from Transformer decoders. Table 2 summarizes loss families by representational style. In the discrete setting, visual tokens are trained with the same cross-entropy objective as text. In the continuous setting, a range of objectives is used. Early approaches regress latent features with cosine or mean squared error losses, as in SEED and EMU2, but these objectives struggle to capture multimodal visual distributions, often yielding over-smoothed images with limited diversity, an issue highlighted by GIVT (Tschannen et al., 2024). To address this limitation, MAR (Li et al., 2024b) adopts a diffusion-style objective that can model arbitrary distributions and has since become common in unified MLLMs, alongside rectified flow based and flow matching based variants. We note, however, that introducing additional objectives increases engineering complexity and raises nontrivial questions about how to weight and balance losses across modalities and tasks. Table 2 provide detailed categories for different loss functions in the framework of unified MLLMs.

3.1.5 Visual Decoder

Finally, we consider the visual decoder in unified MLLMs. VAE-based encoders naturally pair with decoders that reconstruct images from latent representations, whereas CLIP-style encoders lack native decoders due to their contrastive training. To bridge this gap, EMU (Sun et al., 2023) finetunes a diffusion model as a decoder by conditioning it on CLIP features, enabling high-fidelity reconstruction while leveraging pretrained diffusion weights. Although decoders are often trained separately from the MLLM, directly conditioning diffusion models on MLLM hidden states has shown promise for tighter coupling and improved generation quality, as shown in Qwen-Image (Wu et al., 2025a) and OmniGen2 (Wu et al., 2025c).

Vision Type	Loss	Generation Diversity	Methods
Discrete	CE-Loss	✓	Chameleon, UniGen,Janus, Janus-Pro, EMU3, MoMa, Manzano
Continuous	Rectified Flow	✓	BEGAL, Janus-Flow
	Flow Matching	✓	Mogao, Show-o2
	Diffusion Loss	✓	UniPic2, MetaQuery, OpenUni, Transfusion,UniFluid, UniPic
	MSE	✗	EMU2, EMU, SEED-X, SEED
	Cosine	✗	MetaMorph
	MSE + Cosine	✗	Nexus-Gen

Table 2: Different losses applied to visual tokens in the framework of unified MLLMs.

3.1.6 Beyond Text and Vision

While most unified MLLMs focus on text and vision modalities, as demonstrated in Table 1, unified MLLMs are not limited to text and vision, other modalities like audio, action, 3D point clouds, etc., also exhibit promising achievements.

Audio. Audio extends unified MLLMs from seeing to listening and speaking. Representative models include MIO(Wang et al., 2024d) and Qwen2.5-Omni (Xu et al., 2025), which incorporate audio alongside text and images, and more recent omni-style systems such as Ming-Omni (AI et al., 2025) and JavisGPT (Liu et al., 2025a) that combine audio with other modalities under a unified framework. Compared to text-vision, audio introduces stronger temporal structure and synchronization constraints, which makes multi-turn interaction and cross-modal alignment a central challenge.

Video. Video pushes unified MLLMs toward temporal reasoning and dynamic generation. Early unified systems such as EMU2 (Sun et al., 2024b) and subsequent EMU3 (Wang et al., 2024c) explicitly include video generation as part of the unified capability, and more recent work further treats video as a primary target for generation (Liu et al., 2025a).

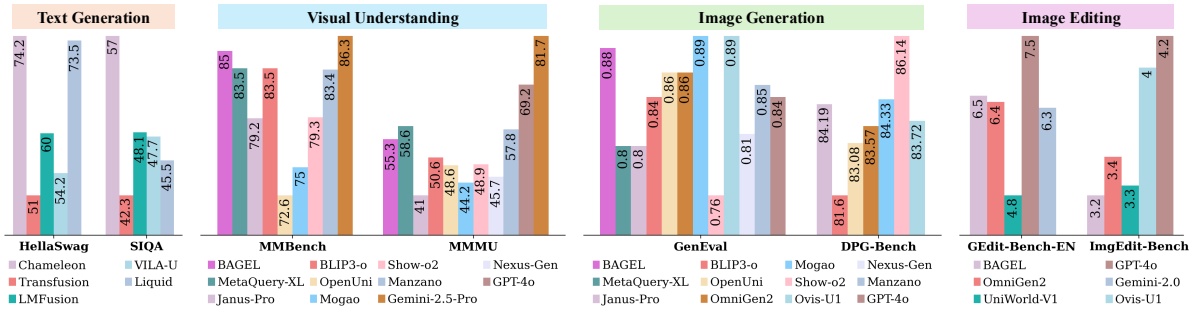

Figure 2: Performance of existing Unified MLLMs across various tasks within text and vision modalities. We report the common-used evaluation metrics for text generation, visual understanding, visual generation, and visual editing.

Beyond images, video amplifies data and modeling challenges due to longer sequences, temporal coherence requirements, and the need to maintain consistency across frames, making it a natural stress test for unified objectives and training stability.

Action and agentic modalities. Action-centric unified models aim to couple multimodal perception with decision-making signals, enabling agent behaviors rather than only content generation. Magma (Yang et al., 2025) is a representative effort in this direction, framing multimodal agents within a unified foundation model perspective. Such settings broaden the notion of unification from generation to interaction, where outputs may correspond to action plans, tool-use decisions, or control commands, and evaluation is less standardized than in text–vision benchmarks. We believe agent-centric unified models would be a great interest in the future research directions.

3D modalities. 3D introduces geometric structure representation, often requiring generation in formats distinct from pixels. Hunyuan3D-Omni (Hunyuan3D et al., 2025) extends unified generation to 3D representations including voxel and point, demonstrating that unified modeling can be applied to spatial modalities when appropriate supervision and decoders exist. Importantly, 3D modalities highlight challenges in data availability, evaluation protocols, and cross-modal grounding between text prompts, 2D inputs and 3D outputs.

Scientific and structured modalities. Unified MLLMs have also demonstrated great potential in scientific modalities like RNA and Protein, and structured modalities. ProLLaMA (Lv et al., 2025) is an example that treats protein generation under the unified modeling framework via next-token-prediction, pointing to a broader direction where unified MLLMs may serve as general-purpose tool for scientific modalities.

3.2 Tasks and Evaluations

3.2.1 Tasks

In contrast to conventional LLMs, which primarily focus on text generation, and VLLMs, which are tailored for visual understanding, unified MLLMs pursue a single framework capable of addressing a wide range of multimodal tasks across diverse domains. These encompass text generation, visual comprehension, audio analysis, visual synthesis, image editing, interleaved text–image generation, video production, audio generation, and beyond. The core premise is that a unified MLLM can seamlessly process multimodal inputs and produce multimodal outputs, as illustrated in Fig. 1. In practice, however, most existing research remains centered on text–vision modalities, with unified approaches for visual understanding and generation emerging as the most actively investigated directions.

3.2.2 Evaluations

Task-Specific Evaluations As mentioned above, unified MLLMs encompass a wide range of modalities and domain-specific tasks. A common evaluation practice is to assess each capability *independently*. Following this principle, existing studies adopt well-established benchmarks tailored to individual tasks. For text generation, representative evaluations include PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2020); for visual understanding, commonly used benchmarks are VQAv2 (Goyal et al., 2017), MMBench (Liu et al., 2024d), SEED (Li et al., 2023a), MM-Vet (Yu et al., 2023c), MMMU (Yue et al., 2024), and TEXTVQA (Singh et al., 2019); for image generation, typical metrics include GenEval (Ghosh et al., 2023), VQAscore (Lin et al., 2024b), WISE (Niu et al., 2025), and DPG-Bench (Hu et al., 2024);

and for image editing, evaluations such as Emu-Edit (Sheynin et al., 2024), GEdit-Bench-EN (Liu et al., 2025b), and ImgEdit-Bench (Ye et al., 2025) are widely adopted. In addition, several benchmarks and evaluation protocols are inherited from their respective domains to enable focused, standalone assessments. We provide detailed results across different tasks and domains in Fig. 2.

Unified Evaluations Recently, increasing attention has been given to evaluating unified MLLMs *holistically*, instead of assessing each modality in isolation. A representative work in this direction is UniEval (Li et al., 2025b), which introduces UniBench alongside the corresponding UniScore metric to jointly measure multimodal understanding and generation capabilities. Similarly, RealUnify (Shi et al., 2025) proposes a benchmark specifically designed to assess bidirectional synergy between visual understanding and generation, thereby examining whether unified MLLMs truly enable cooperative interactions among constituent modalities. We anticipate the emergence of more such holistic evaluation protocols that move beyond task-specific pipelines toward genuinely integrated multimodal assessment.

3.3 Training Strategy and Data mixing

3.3.1 Training strategy

Unified MLLMs typically train from scratch, Unified MLLMs always initialize with pretrained backbones (e.g., LLMs, VLMs) and consider modality interfaces such as projectors or adapters (Chen et al., 2025b; Pan et al., 2025) for further continue training. This staged approach is pragmatic, it mitigates optimization instability and prevents the regression of pretrained language capabilities that can happen when early multimodal gradients are always noisy (Team, 2024). Another core design is the multimodal generation mechanism. Discrete approaches unify generation via next-token prediction over a shared vocabulary (Team, 2024; Liu et al., 2024a), whereas continuous approaches often decouple the LLM from other modalities' generation (Chen et al., 2025b). As a result, many SOTA systems adopt hybrid recipes, stabilizing the unified backbone before integrating generative decoders to ensure high-fidelity output without destabilizing the core model.

3.3.2 Data mixing

Data mixing in unified MLLMs balances different modality-aware and task-aware data sources (Deng et al., 2025; Liu et al., 2024a). This follows an implicit experience via stage-specific ratios: bootstrapping alignment with paired data, expanding to interleaved context for compositional generalization, and concluding with instruction-tuning for controllability. Crucially, the mixing ratio is always measured in *tokens*, in case of small fraction of multimodal samples may dominate compute and gradient updates. As a result, mixture design usually couples *sampling ratios* with *token-budget constraints*, and both are adjusted across stages to maintain stable optimization while gradually expanding modality and task coverage.

4 Challenges and Outlook

4.1 Challenges

Efficiency Training unified MLLMs to reasonable performance typically demands large parameter counts and substantial compute, as shown by Chameleon (Team, 2024): training a 34B model required 4,282,407 GPU hours on 3,072 GPUs—well beyond most research groups' reach—and its performance still trails state-of-the-art standalone models. A pragmatic path to efficiency is to initialize from strong VLMs or LLMs, leveraging pretrained weights and world knowledge; works such as BLIP3-o and LMFusion report notable gains under this strategy. In parallel, sparse attention and hardware-aligned designs (Yuan et al., 2025) show promising potential.

Current dilemma Beyond efficiency, unified MLLMs remain dilemma in grounding: they can hallucinate unexpected details (Kalai et al., 2025; Xu et al., 2024b) and show weak reasoning and limited real-world modeling (Hong et al., 2025). Recent self-supervised video learning offers a concrete path toward richer temporal world priors. for example, V-JEPA 2 (Assran et al., 2025) learns predictive video representations at scale and improves motion understanding for downstream tasks after LLM alignment. Integrating such video-pretrained dynamics into unified MLLMs may help reduce hallucinations and push toward more world-aware multimodal intelligence.

Performance Despite steady progress, unified MLLMs have not yet surpassed standalone models tailored to a single task or modality. Moreover, at matched parameter counts, they often underperform comparable LLMs or VLMs, showning a persistent performance regression. Closing this gap

remains a priority. As suggested by Manzano (Li et al., 2025a) and related work, a central requirement is sufficient model capacity and and corresponding training signals to competently handle multiple modalities and task families.

Large-scale implementation Most critically, unified MLLMs demand large model capacity and substantial training data, imposing stringent requirements for large-scale deployment. Efficient parallelization and distributed training are therefore indispensable, and any extra modules or bespoke components can add significant implementation burden across the pipeline. We noticed that Token-Shuffle (Ma et al., 2025b) provided detailed insights about this. We thus advocate for simple, compute-aware designs that preserve efficiency while advancing the goals of unified MLLMs.

4.2 Outlook

Although numerous unified MLLMs have been proposed and analyzed architecturally, their underlying motivations and practical benefits remain insufficiently articulated. Clarifying these drivers is essential for advancing the field. In this section, we synthesize the principal motivations and concrete benefits of unification to inform both research directions and deployment practice.

Unified MLLMs are a solution to the Platonic representation by learning a shared statistical model of reality. Different modalities provide biased views of the same world, yielding partial and modality-specific representations. A central question is how to learn the underlying, modality-agnostic structure. Inspired by platonic representation, we posit that unified MLLMs offer a practical route to approximate this shared latent representation and its statistics by jointly modeling multimodal representations across modalities.

Unified MLLMs are key to mutual benefits across modalities Training each modality in isolation prevents information and inductive biases from being shared across modalities. A unified, shared platonic representation, as discussed above, offers a more general abstraction that can support consistent reasoning and transfer across diverse input types. Despite many proposed unified MLLMs, strong and reproducible evidence of mutual benefits across modalities remains limited. The shortfall likely reflects constraints in model capacity, data scale and balance, and overall computational budget. Closing this gap is central to the promise of unified MLLMs: with sufficient capacity and appropriately curated multimodal data, unified MLLMs should enable knowledge learned in one modality to improve learning and generalization in another.

Unified MLLMs enable mutual benefits across tasks within the same modality. Unified MLLMs aim not only to bridge modalities but also to consolidate heterogeneous tasks within a single modality. A canonical case is vision, where image generation and visual understanding are treated within one framework. Although some systems employ separate encoders for the two tasks, recent studies indicate that a single shared visual encoder can yield mutual benefits by aligning representations and training signals across tasks, as evidenced by MetaMorph (Tong et al., 2024) and the results reported in (Zhang et al., 2025b).

Unified MLLMs can be zero-shot task learners. Beyond cross-modal and within-modal representation learning, we further hypothesize that unified MLLMs can act as zero-shot task learners within a modality. Concretely, even when training is limited to mixed text data, image understanding data, and image generation data, a unified MLLM may spontaneously induce related capabilities such as image editing and interleaved text–image generation. While performance on these unseen tasks may lag behind specialized systems, the emergence of such zero-shot behaviors is noteworthy: it suggests a path toward more general intelligence, where enumerating every potential task is neither necessary nor feasible. Continued progress will likely depend on scaling capacity, broadening the diversity of training signals, and designing objectives that encourage transferable, compositional skills rather than task-specific shortcuts.

5 Conclusion

In this survey, we review recent advances in unified multimodal LLMs, organize the literature with a coherent methodological taxonomy, and discuss detailed design guidance. We situate unified models within their background and enabling technologies, identify the key challenges that limit current systems, and outline concrete directions for future work. We hope our survey can offer a clear reference that clarifies recent advances in the filed of unified MLLMs, and help both researchers and practitioners deepen understanding and accelerate progress on unified MLLMs.

Limitations

The fast development of progress in unified MLLMs makes it impractical to cover every new technique. Instead, we focus on a set of recent, representative methods and offer clear, detailed analyses of their designs and trade-offs. Many leading systems (e.g., Gemini and ChatGPT models) remain closed, often without code and with limited technical documentation. Hence, deep and verifiable insight into those models is not always possible. When direct details are unavailable, we make cautious, evidence-based inferences and anchor our analysis in open, reproducible implementations. Finally, we highlight that detailed training strategy and training data are decisive for performance and often matter as much as architectural choices. However, these detailed are always missed in state-of-the-art models.

Ethics Statement

We confirm that this survey adheres to established research ethics and community norms. Our analysis draws exclusively on publicly available papers, datasets, and implementations; it involves no human or animal subjects, confidential materials, or personal data. All works are properly cited, and we have made efforts to present prior work accurately and in context. Where we reference code or datasets, we respect their licenses and acknowledge the original creators. The interpretations and synthesis presented are our own original contributions. The authors report no conflicts of interest. Besides, we used AI tools only for language editing and minor formatting; they did not generate ideas, analyses, results, figures, or layouts. All content and design are original to the authors.

References

Gemini 2.5 flash image (nano banana) — google ai studio. https://aistudio.google.com/models/gemini-2-5-flash-image. Accessed: 2025-10-05.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, and 1 others. 2025. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, and 1 others. 2025. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, and 1 others. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. 2025. Flextok: Resampling images into 1d token sequences of flexible length. In *Forty-second International Conference on Machine Learning*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng Zhang, Tong He, and Mike Zheng Shou. 2024. Factorized visual tokenization and generation. *arXiv preprint arXiv:2411.16681*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, and 1 others. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, and 1 others. 2025. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Fengjiao Chen, Minhao Jing, Weitao Lu, Yan Feng, Xiaoyu Li, and Xuezhi Cao. 2025a. Unihetero: Could generation enhance understanding for vision-language-model at large data scale? *arXiv preprint arXiv:2512.23512*.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, and 1 others. 2025b. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR.

Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. 2024. Eat: Self-supervised pre-training with efficient audio transformer. *arXiv preprint arXiv:2401.03497*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025c. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, and 1 others. 2025. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, and 1 others. 2025. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. 2025. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744.

Philippe Hansen-Estruch, David Yan, Ching-Yao Chung, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. 2025. Learnings from scaling visual tokenizers for reconstruction and generation. *arXiv preprint arXiv:2501.09755*.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. 2025. Uni-x: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. *arXiv preprint arXiv:2509.24365*.

Xin He, Longhui Wei, Jianbo Ouyang, Lingxi Xie, and Qi Tian. 2025. Emma: Efficient multimodal understanding, generation, and editing with a unified architecture. *arXiv preprint arXiv:2512.04810*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Team Hunyuan3D, Bowen Zhang, Chunchao Guo, Haolin Liu, Hongyu Yan, Huiwen Shi, Jingwei Huang, Junlin Yu, Kunhong Li, Penghao Wang, and 1 others. 2025. Hunyuan3d-omni: A unified framework for controllable generation of 3d assets. *arXiv preprint arXiv:2509.21245*.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jiyue Jiang, Zikang Wang, Yuheng Shan, Heyan Chai, Jiayi Li, Zixian Ma, Xinrui Zhang, and Yu Li. 2025a. Biological sequence with language model prompting: A survey. *arXiv preprint arXiv:2503.04135*.

Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. 2025b. From specific-mllms to omni-mllms: a survey on mllms aligned with multimodalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8617–8652.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Jinwoo Kim, Dat Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. 2022. Pure transformers are powerful graph learners. *Advances in Neural Information Processing Systems*, 35:14582–14595.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image

generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024b. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445.

Yanghao Li, Rui Qian, Bowen Pan, Haotian Zhang, Haoshuo Huang, Bowen Zhang, Jialing Tong, Haoxuan You, Xianzhi Du, Zhe Gan, and 1 others. 2025a. Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer. *arXiv preprint arXiv:2509.16197*.

Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. 2025b. Unieval: Unified holistic evaluation for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.10483*.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. 2025. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. 2024a. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and 1 others. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024b. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. 2024a. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Jianzhang Gao, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, and 1 others. 2025a. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. *arXiv preprint arXiv:2512.22905*.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, and 1 others. 2025b. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024d. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, and 1 others. 2025c. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787.

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. 2025. Atoken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476*.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2025. Prollama: A protein large language model for multi-task protein language processing. *IEEE Transactions on Artificial Intelligence*.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. 2025a. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*.

Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. 2022. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*.

Xu Ma, Peize Sun, Haoyu Ma, Hao Tang, Chih-Yao Ma, Jialiang Wang, Kunpeng Li, Xiaoliang Dai, Yujun Shi, Xuan Ju, and 1 others. 2025b. Token-shuffle: Towards high-resolution image generation with autoregressive models. *arXiv preprint arXiv:2504.17789*.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, and 1 others. 2025c. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sicheng Mo, Thao Nguyen, Xun Huang, Siddharth Srinivasan Iyer, Yijun Li, Yuchen Liu, Abhishek Tandon, Eli Shechtman, Krishna Kumar Singh, Yong Jae Lee, and 1 others. 2025. X-fusion: Introducing new modality to frozen large language models. *arXiv preprint arXiv:2504.20996*.

Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, and 1 others. 2025. Discrete audio tokens: More than a survey! *arXiv preprint arXiv:2506.10274*.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, and 1 others. 2025. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*.

Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason E Weston, Luke Zettlemoyer, and 1 others. 2025. Byte latent transformer: Patches scale better than tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9238–9258.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, and 1 others. 2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*.

Yatian Pang, Eng Hock Francis Tay, Li Yuan, and Zhenghua Chen. 2023. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1:2440001.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. 2025. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2024. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. 2024. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, and 1 others. 2025. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024a. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Sabera Talukder, Yisong Yue, and Georgia Gkioxari. 2024. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2024. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. 2024. Givt: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision*, pages 292–309. Springer.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, and 1 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, and 1 others. 2025a. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning. *arXiv preprint arXiv:2505.07538*.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, and 1 others. 2025b. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*.

Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. 2024a. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295.

Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, and 1 others. 2025c. Skywork unipic: Unified autoregressive modeling for visual understanding and generation. *arXiv preprint arXiv:2508.03320*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024c. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, and 1 others. 2024d. Mio: A foundation model on multimodal tokens. *arXiv preprint arXiv:2409.17692*.

Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, and 1 others. 2025. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025a. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and 1 others. 2025b. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, and 1 others. 2025c. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. 2024a. Liquid: Language models are scalable and unified multi-modal generators. *arXiv preprint arXiv:2412.04332*.

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, and 1 others. 2024b. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. 2025d. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. 2024c. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, and 1 others. 2024d. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and 1 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*.

Yi Xin, Juncheng Yan, Qi Qin, Zhen Li, Dongyang Liu, Shicheng Li, Victor Shea-Jay Huang, Yupeng Zhou, Renrui Zhang, Le Zhuo, and 1 others. 2025. Lumina-mgpt 2.0: Stand-alone autoregressive image modeling. *arXiv preprint arXiv:2507.17801*.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. 2025. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. *arXiv preprint arXiv:2504.08736*.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data. *arXiv preprint arXiv:2309.16671*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yifan Xu, Xiaoshan Yang, Yaguang Song, and Changsheng Xu. 2024a. Libra: Building decoupled vision system on large language models. *arXiv preprint arXiv:2405.10140*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, and 1 others. 2025. Magma: A foundation model for multimodal ai agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14203–14214.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. 2025. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and 1 others. 2023a. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, and 1 others. 2023b. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. 2024. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023c. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, and 1 others. 2025. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. 2025a. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, and 1 others. 2025b. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*.

Yitian Zhang, Long Mai, Aniruddha Mahapatra, David Bourgin, Yicong Hong, Jonah Casebeer, Feng Liu, and Yun Fu. 2025c. Regen: Learning compact video embedding with (re-) generative decoder. *arXiv preprint arXiv:2503.08665*.

Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. 2024a. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*.

Long Zhao, Sanghyun Woo, Ziyu Wan, Yandong Li, Han Zhang, Boqing Gong, Hartwig Adam, Xuhui Jia, and Ting Liu. 2024b. epsilon-vae: Denoising as visual decoding. *arXiv preprint arXiv:2410.04081*.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2023a. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Detailed Visual Tokenizers

While text tokenizers are covered in detail elsewhere in this survey, here we present a finer-grained categorization of visual tokenizers applicable to unified MLLMs, summarized in Table 3.

B Vision-Language Understanding

A central goal of VLMs is to jointly model an image $x_v \in \mathbb{R}^{H\times W\times C}$ and a text sequence $x_t = (w_1, w_2, \ldots, w_n)$, enabling cross-modal reasoning and generation. Early methods such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) optimized a contrastive objective:

$$\mathcal{L}_{\text{CLIP}} = -\sum_i \log \frac{\exp(\langle f_v(x_v^i), f_t(x_t^i)\rangle/\tau)}{\sum_j \exp(\langle f_v(x_v^i), f_t(x_t^j)\rangle/\tau)},$$

where f_v, f_t are image and text encoders, and τ is a temperature. These models demonstrated strong zero-shot transfer, but their reliance on global embeddings limited fine-grained grounding.

To achieve better cross-modal integration, models such as BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023b) introduced multimodal pretraining strategies combining image-text matching and captioning objectives. BLIP-2 further proposed a *Q-former*, a lightweight query transformer, to project visual tokens into the language embedding space of a frozen LLM. This modular paradigm paved the way for scalable integration of pretrained vision encoders and LLMs.

Recent works move toward instruction-following multimodal models. LLaVA (Liu et al., 2023), MiniGPT-4 (Zhu et al., 2023b), and InstructBLIP (Dai et al., 2023) couple a vision encoder with an LLM (e.g., Vicuna, LLaMA) and finetune them with multimodal instruction datasets, effectively learning a mapping $p(y \mid x_v, x_t; \theta)$ where y is an instruction-compliant response. These methods significantly improve performance on visual question answering, captioning, and dialogue-based benchmarks.

Overall, the development of VLMs reflects a progression from *contrastive alignment* of independent encoders, to *modular integration* with frozen LLMs, to *unified autoregressive modeling* where vision and language are treated symmetrically. This trajectory highlights the increasing emphasis on efficiency, scalability, and instruction alignment, laying the foundation for current unified multimodal large language models.

C Image Generation with Diffusion

We introduce one important task in unified MLLM, image generation. We start with the image genera-

Model	Discrete	Continuous	Pixel Level	Semantic Level	Encoder Arch	Decoder Arch	Native Resolution	Spatial Compression
VQ-VAE (Van Den Oord et al., 2017)	✓	✗	✓	✗	Conv	Conv	-	4×
VQ-GAN (Esser et al., 2021)	✓	✗	✓	✗	Conv	Conv	-	16×
SD-VAE (Rombach et al., 2022)	✗	✓	✓	✗	Conv	Conv	✓	8×
FQ-GAN (Bai et al., 2024)	✓	✗	✓	✗	Conv	Conv	-	8/16×
GigaTok (Xiong et al., 2025)	✓	✗	✓	✗	Hybrid	Hybrid	✗	16×
TiTok (Yu et al., 2024)	✓	✗	✗	✓	Tran	Tran	-	∼32×
FlexTok (Bachmann et al., 2025)	✓	✗	✗	✓	Tran	Tran (Diff)	-	∼32×
Selftok (Wang et al., 2025a)	✓	✗	✗	✓	Tran	Tran (Diff)	-	∼8×
ϵ-VAE (Zhao et al., 2024b)	✗	✓	✓	✗	Conv	Conv (Diff)	-	8/16×
REGEN (Zhang et al., 2025c)	✗	✓	✓	✗	Conv	Tran (Diff)	✓	8×
OmniTokenizer (Wang et al., 2024a)	✓	✓	✓	✗	Tran	Tran	✗	8×
MAGVIT-v2 (Yu et al., 2023b)	✓	✗	✓	✗	Conv	Conv	-	8×
Cosmos (Agarwal et al., 2025)	✓	✓	✓	✗	Conv	Conv	-	8/16×
ViTok (Hansen-Estruch et al., 2025)	✗	✓	✓	✗	Tran	Tran	✗	16×
HunyuanImage3.0 (Cao et al., 2025)	✗	✓	✓	✗	Conv	Conv	✓	16×
Wan2.1 (Wan et al., 2025)	✗	✓	✓	✗	Conv	Conv	✓	8×
VILA-U (Wu et al., 2024d)	✓	✗	✓	✓	Tran	Conv	✗	16×
TokenFlow (Qu et al., 2025)	✓	✗	✓	✓	Tran	Tran	-	16×
UniTok (Ma et al., 2025a)	✓	✗	✓	✓	Tran	Hybrid	✗	16×
AToken (Lu et al., 2025)	✓	✓	✓	✓	Tran	Tran	✓	16×

Table 3: Comparison of existing visual tokenizers. We summarize and contrast representative visual tokenizers across various dimensions. 'Diff' refers to diffusion based generative decoder and '∼' represents the approximate spatial compression rate based on the number of latent tokens.

tion with Diffusion-based model. Diffusion models have become the dominant paradigm for image generation by modeling data distributions through the reversal of a gradual noising process. The forward process adds Gaussian noise to a clean image x_0:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_0, (1-\alpha_t) I), \quad (1)$$

while the model learns the reverse distribution by predicting noise ϵ, optimized via

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \quad (2)$$

This formulation, realized in DDPM (Ho et al., 2020), Imagen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022), produces high-fidelity images, especially under text conditioning, but requires hundreds of iterative denoising steps.

Subsequent work re-framed diffusion as continuous-time flows to improve efficiency. Deterministic solvers such as DDIM (Song et al., 2020) and DPM-Solver (Lu et al., 2022) reduced sampling steps by interpreting the process as an ODE. Flow matching (Lipman et al., 2022) generalizes this idea by directly learning a velocity field $v_\theta(x_t, t)$ that transports noise to data via

$$\frac{dx_t}{dt} = v_\theta(x_t, t). \quad (3)$$

This deterministic formulation avoids stochastic denoising, achieves high-quality samples in tens of steps, and unifies diffusion with normalizing flows. Recent methods such as Rectified Flow (Liu et al., 2022) demonstrate both efficiency and fidelity, establishing flow-based models as promising backbones for multimodal generative systems.

D Image Generation with Autoregressive Models

Unlike Diffusion models, Autoregressive (AR) image generation extends the next-token prediction paradigm of language modeling to discrete visual sequences. An image I is mapped to tokens $x = (x_1, \ldots, x_T)$ via a learned tokenizer (e.g., DVAE/VQ-VAE, RQ-VAE (Lee et al., 2022), VQ-GAN (Esser et al., 2021)), and the model maximizes the chain-rule likelihood

$$P(x) = \prod_{t=1}^{T} P(x_t \mid x_{<t}; \theta), \quad (4)$$

typically with a Transformer. This formulation inherits the scaling behavior and controllability of LLMs while shifting the core design choices to *1)* tokenization (continuous or discrete), *2)* decoding strategy (de-nosing or token-prediction), and *2)* generation order (global or causal).

Historically, pixel/patch AR models (Image Transformer, iGPT (Chen et al., 2020)) demonstrated viability but were compute-heavy. Discrete latents made AR practical at scale:

DALL·E (Ramesh et al., 2021) showed compositional text-to-image with a DVAE prior; RQ-Transformer (Lee et al., 2022) improved fidelity via hierarchical/residual codes; MAGVIT (Yu et al., 2023a) and MAGVIT-v2 (Yu et al., 2023b) unified efficient tokenizers for images and video. Two influential AR decoding families emerged: *umasked-parallel infilling* (e.g., MaskGIT), which iteratively predicts subsets $x_{\mathcal{M}} \sim P(x_{\mathcal{M}} \mid x_{\setminus\mathcal{M}})$ to accelerate sampling and allow bidirectional context; and causal next-token-prediction (e.g., LlamaGen (Sun et al., 2024a), Parti (Yu et al., 2022)), which directly ports LLM training and sampling to visual tokens, yielding stable optimization, clean conditioning, and strong scaling.

A complementary axis rethinks generation order: hierarchical and coarse-to-fine schemes (e.g., VAR (Tian et al., 2024) and variants) predict low-resolution or low-frequency structure first, then refine details, narrowing the effective context length and improving throughput; hybrid designs add continuous residuals or multi-scale priors to close the high-frequency gap. Multimodal AR systems like Chameleon (Team, 2024) further interleave text and image tokens in one sequence, suggesting a unified AR interface for composition and editing.

AR image generation is best viewed as LLM-style modeling over discrete visual codes: tokenizers set the information bottleneck; masked AR trades a bit of global consistency for speed via parallel refinement; causal AR trades speed for simplicity, controllability, and scaling laws; hierarchical ordering (coarse-to-fine) reconciles both by shortening dependency paths. Together, these threads have pushed AR methods to parity with diffusion on quality while offering cleaner conditioning, easier reuse of LLM infrastructure, and the promise of integration with LLM for a unified multimodal framework.

E VLM Performance Preservation

As expected, building unified MLLMs upon pretrained VLMs or LLMs necessitates the introduction of additional capabilities. Hence, a common practice is to initialize unified models with pretrained VLMs & LLMs weights. However, this often leads to performance regression on core tasks such as text generation and visual understanding. To mitigate this issue and facilitate efficient training, several approaches adopt a strategy of freezing the pretrained LLM or VLM while introducing generative capabilities through plug-and-play generation modules, as exemplified by BLIP3-o (Chen et al., 2025b), MetaQuery (Pan et al., 2025), and OpenUni (Wu et al., 2025d). Similarly, the recent Nano-Banana model (gem) is expected to follow this paradigm, extending Gemini with promising image generation capability. While such designs appear simple and effective, we argue that they merely emulate unified MLLMs rather than embody their ultimate form, as they compromise the foundational motivation of true unified MLLMs discussed in Sec. 4.2.