Unifying Understanding and Generation in Vision-Language Models: Advances, Challenges, and Opportunities

Anonymous authors
Paper under double-blind review

Abstract

Significant advancements in vision-language models have predominantly followed two divergent trajectories: autoregressive architectures optimized for visual understanding and diffusion-based frameworks designed for high-fidelity generation. However, this separation hinders the development of truly versatile multimodal agents. Unifying these capabilities is a critical step toward Artificial General Intelligence, as recent findings suggest that effective understanding and generation can mutually reinforce each other. This survey provides a comprehensive overview of the emerging field of unified vision-language models and proposes a systematic taxonomy based on the core visual representation mechanism: *continuous* versus *discrete* visual tokens. For continuous visual tokens, we analyze how models bridge the semantic-visual gap by categorizing integration strategies into Serial Coupling, where LLMs act as planners, and Parallel Coupling, which enables bidirectional interaction. regarding discrete visual tokens, we contrast Autoregressive approaches that treat images as a foreign language against emerging Discrete Diffusion paradigms known for their global consistency and parallel decoding. Beyond architectural analysis, we provide a curated compilation of datasets and benchmarks essential for training and evaluation. Finally, we critically discuss open challenges such as tokenization trade-offs, training stability, and scalability, while outlining future directions for building seamless, omni-capable multimodal systems.

1 Introduction

The past few years have witnessed a transformative era in Artificial Intelligence, primarily fueled by the rapid advancement of Large Language Models (LLMs). Models such as LLaMATouvron et al. (2023), PanGu Zeng et al. (2021), Qwen Bai et al. (2023); Team et al. (2024); Yang et al. (2025a), and the GPT series Brown et al. (2020); Achiam et al. (2023) have demonstrated unprecedented capabilities in understanding, reasoning, and text generation, fundamentally reshaping diverse applications. This success has naturally extended to multimodal domains, recognizing that true intelligence necessitates processing and synthesizing information beyond text. This evolution has given rise to powerful Vision-Language Models (VLMs) and advanced visual generation technologies, each progressing along distinct, yet highly impactful, trajectories.

In multimodal understanding, models primarily leverage autoregressive (AR) architectures (As shown in Figure 2a), building on the success of LLMs, to interpret diverse visual and textual inputs and generate discrete textual outputs that showcase deep semantic comprehension and reasoning capabilities. This approach benefits from AR models' strong capabilities in sequential modeling and nuanced reasoning. Conversely, visual generation has undergone its own rapid evolution. While initially explored by Generative Adversarial Networks (GANs) Goodfellow et al. (2020), the field is now overwhelmingly dominated by diffusion models Ho et al. (2020) (As shown in Figure 2b). Architectures like UNet Ronneberger et al. (2015) and Diffusion Transformers (DiT) Peebles & Xie (2023), often conditioned by advanced text encoders such as CLIP Radford et al. (2021) and T5 Raffel et al. (2020), are proficient in synthesizing continuous, high-fidelity visual content from diverse prompts (e.g., Stable Diffusion series Rombach et al. (2022); Podell et al. (2023)). However, pure diffusion-based generation models, despite their photorealism, fundamentally lack the deep world knowledge or true semantic understanding inherent in LLM-based systems. Their outputs are often faithful to the prompt's surface-level descriptions rather than reflecting nuanced conceptual comprehension.

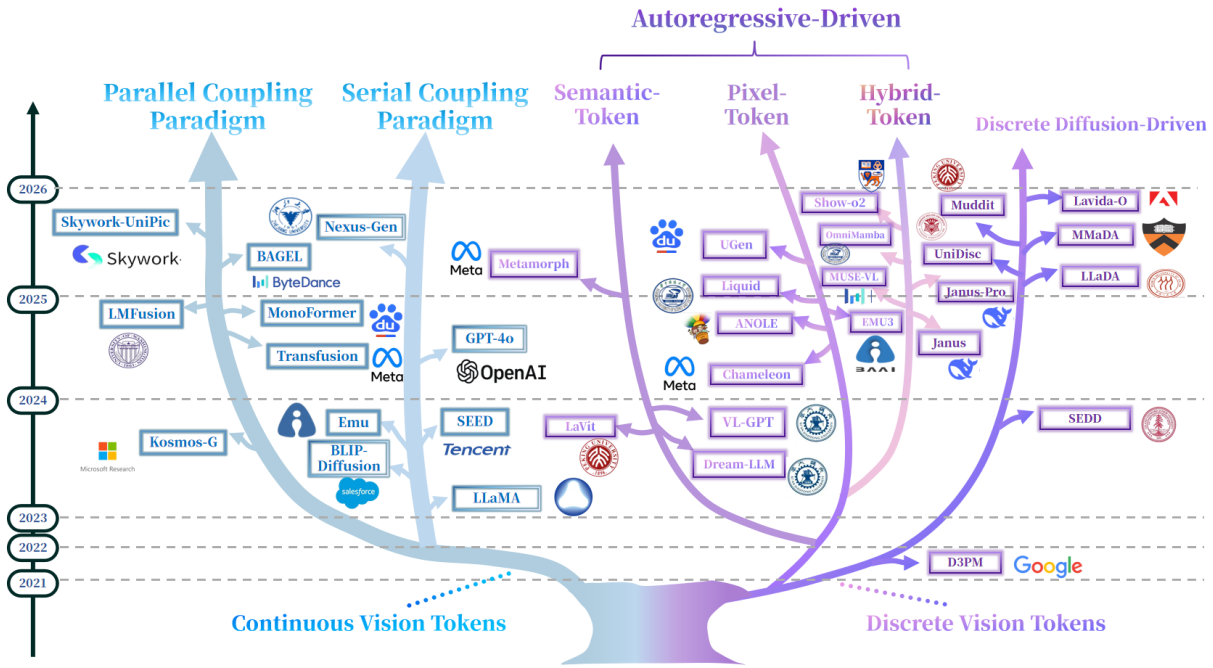

Figure 1: **Evolutionary Timeline (2021–2025) of Unified Understanding–Generation Models.** The diagram traces the convergence of multimodal AI from separate paradigms to unified frameworks. The branches distinguish models by their core visual representation: the blue stream represents *Continuous Vision Tokens* (prioritizing fidelity via diffusion), while the purple stream represent *Discrete Vision Tokens* (prioritizing alignment via AR or Discrete Diffusion). Key milestones highlight the rapid progression toward omni-capable agents.

While some explorations have attempted to adapt LLM-inspired AR architectures for image generation Sun et al. (2024); Tian et al. (2024), diffusion-based approaches currently maintain state-of-the-art performance in terms of visual quality.

The fundamental divergence in these underlying modeling paradigms—where AR is optimized for understanding and discrete text generation, versus Diffusion which excels in continuous visual synthesis but without intrinsic world knowledge—has historically posed a significant challenge for seamless unification into a singular framework. However, the vision of a unified model capable of both understanding and generating multimodal content holds immense potential. Such a framework could revolutionize how humans interact with AI, enabling complex visual reasoning guided by natural language, generating images from abstract instructions, and facilitating dynamic multimodal dialogues. The recent unveiling of GPT-4o Hurst et al. (2024), demonstrating a singular model's ability to profoundly understand complex instructions and generate highly controllable multimodal content across these paradigms, vividly highlights this transformative potential and has galvanized research in this direction.

Designing such a truly unified framework presents substantial technical hurdles. A core challenge lies in the effective integration of AR's robust reasoning and text generation capabilities with Diffusion's generative quality and control. This often boils down to fundamental questions surrounding image representation and processing for unified models. For instance, how should visual information be tokenized for an AR backbone when a continuous latent space is preferred for high-quality generation? Existing approaches vary, from employing VQ-VAEVan Den Oord et al. (2017) or VQ-GAN Esser et al. (2021) to create discrete visual tokens, to utilizing semantic encoders like EVA-CLIPSun et al. (2023a) or OpenAI-CLIP Radford et al. (2021) that produce continuous embeddings. Furthermore, architectural design itself, encompassing purely autoregressive, diffusion-centric, or hybrid structures, presents various trade-offs that are still under active investigation.

To provide a comprehensive overview and foster future research in this rapidly converging and critical field (As shown in Figure 1), this survey systematically analyzes unified multimodal understanding and generation models. Our unique contribution lies in classifying existing unified models based on their core visual representation mechanisms: discrete visual tokens and continuous visual tokens. This classification enables a clearer comparison between models that emphasize the integration of textual understanding with discrete visual tokens (e.g., autoregressive models) and those that leverage continuous visual representations (e.g., diffusion models) for high-fidelity image generation. By highlighting the trade-offs and synergies between these approaches, we aim to uncover the strengths and limitations of each mechanism in the context of multimodal unification. Moreover, we discuss the implications of these design choices on model efficiency, scalability, and the ability to generate semantically rich and coherent visual content. Through this structured analysis, we aim to provide valuable insights and a roadmap for researchers seeking to advance the unification of understanding and generation within vision-language models, ultimately contributing to more powerful, efficient, and flexible multimodal AI systems.

Taxonomy Preview Unlike previous reviews that categorize models broadly by architecture, this work offers a specialized analysis grounded in the fundamental strategy of visual representation. As illustrated in Figure 1, we propose a hierarchical taxonomy that divides Unified Vision-Language Models into two primary categories, each with distinct sub-paradigms that offer unique trade-offs between flexibility, fidelity, and coherence.

Unified Models with Continuous Vision Tokens. These models interface LLMs with continuous latent representations, typically leveraging diffusion models to prioritize high-fidelity generation. We classify them based on how the reasoning (LLM) and generation (Diffusion) modules interact:

- **Serial Coupling Paradigm:** In this approach, the LLM functions as a "Semantic Planner," converting instructions into static embeddings that condition a separate visual generator (e.g., Emu, SEED). *Preview:* This paradigm excels in **modularity and scalability**, allowing the flexible combination of state-of-the-art LLMs and diffusion models. However, the unidirectional information flow creates a **"semantic-visual gap,"** as the LLM receives no feedback during the generation process, often limiting fine-grained alignment in complex spatial tasks.
- **Parallel Coupling Paradigm:** This emerging strategy fuses the LLM and visual generator within a unified attention span, enabling bidirectional interaction at each step (e.g., Transfusion, Skywork-UniPic). *Preview:* By allowing text and visual tokens to attend to each other, this paradigm achieves superior **semantic coherence and real-time interleaving**. The primary trade-off is higher **computational complexity** and the need for sophisticated training strategies to synchronize the discrete text space with the continuous latent space.

Unified Models with Discrete Vision Tokens. These models quantize visual data into discrete tokens (via VQ-VAE or VQ-GAN) that share a categorical vocabulary with text, effectively treating images as a "foreign language." We categorize them by their generative mechanism:

- **Autoregressive-Driven (AR):** Following the success of GPT, these models model visual tokens via next-token prediction (e.g., LLaViT, Show-o2). *Preview:* AR models benefit from a **unified training objective** that naturally aligns with LLMs, facilitating strong logical reasoning and instruction following. However, the unidirectional causal attention is suboptimal for visual data (which is naturally bidirectional), leading to issues like **error accumulation** and slow, serial inference speeds for high-resolution images.
- **Discrete Diffusion-Driven:** An innovative paradigm that applies diffusion processes directly to discrete tokens, predicting all tokens in parallel via iterative refinement (e.g., LLaDA, UniDisc). *Preview:* This approach combines the best of both worlds: the **alignment benefits** of a shared vocabulary and the **global coherence** of diffusion. It overcomes the serial bottleneck of AR models, enabling faster **parallel decoding** and self-correction. The challenge lies in the novelty of the method, which currently lacks the mature training recipes and stability of established AR or continuous diffusion frameworks.

Scope and objectives In comparison to existing surveys that broadly cover multimodal understanding or image generation, our work offers a unique and systematic analysis focusing specifically on unified multimodal understanding and generation models. Unlike previous reviews that might categorize models by architectural components (e.g., *autoregressive*, *diffusion*, *hybrid*), our distinctive approach is to classify these unified frameworks based on their core visual generation mechanisms: discrete visual generation and continuous visual generation. This novel categorization allows for a deeper exploration into how models bridge the gap between LLM-based understanding and diverse visual synthesis strategies. We meticulously detail the structural designs, visual representation strategies, and integration techniques with language models for each category. Furthermore, we provide a comprehensive compilation of relevant datasets and benchmarks, critically analyze key challenges including effective tokenization and cross-modal attention, and delineate promising future research directions. Our survey aims to provide a fresh perspective and a structured roadmap for researchers navigating the complexities of building truly integrated multimodal AI systems.

Organization The remainder of this survey is organized as follows. Chapter 2 lays the groundwork by introducing the fundamental concepts and recent advances in Large Language Models (LLMs), multimodal understanding, and visual generation, setting the stage for unified models. Chapter 3 focuses on unified models that use Continuous Visual Generation, detailing their unique characteristics and technical innovations. Following this, in Chapter 4, we dive into unified models that employ discrete visual generation mechanisms, analyzing their architectural designs, visual representation strategies, and integration with language models. Chapter 5 provides a comprehensive compilation of relevant datasets and benchmarks crucial to the training and evaluation of these sophisticated unified models. Finally, Chapter 6 discusses the key challenges confronting this nascent field and outlines promising future research directions, aiming to inspire further advancements in unified multimodal AI.

2 Background

This chapter provides an overview of the background behind Vision-Language Models (VLMs), focusing on their unified capabilities in both understanding and generation. We start by reviewing Large Language Models (LLMs), which have revolutionized NLP with their generative and reasoning abilities. We then explore visual understanding methods, detailing how visual information is encoded and integrated with LLMs. Finally, we discuss text-to-image generation models, which, combined with visual understanding, lay the foundation for unified multimodal AI.

2.1 Large Language Models

The Transformer architecture's rise, initially enabling discriminative (e.g., BERT Devlin et al. (2019)) and generative (e.g., GPT-2 Radford et al. (2019)) NLP tasks, culminated in a paradigm shift with generative large language models (LLMs). Driven by exponential scaling and enhanced by techniques like instruction tuning Shengyu et al. (2023) and RLHF Ouyang et al. (2022), LLMs unified diverse NLP applications into sophisticated instruction-following and reasoning frameworks Wei et al. (2022). Their emergent abilities and tool integration have transformed them into versatile agents, establishing LLMs as the crucial foundation for unified multimodal intelligence.

2.2 Vision Representations Methods

Vision representations play a critical role in bridging visual inputs with language models in unified vision-language systems. These representations can be learned using both supervised and unsupervised (or self-supervised) methods, each offering distinct advantages depending on the task and available data. In this section, we provide an overview of the most widely used approaches for visual representations, categorized into supervised and unsupervised learning.

Supervised Learning-Based Methods. In supervised learning, vision representations are learned from labeled datasets, with models trained to map images to feature vectors aligned with specific labels. CNNs and Vision Transformers (ViT) are commonly used as backbones. ResNet He et al. (2016) introduced deep

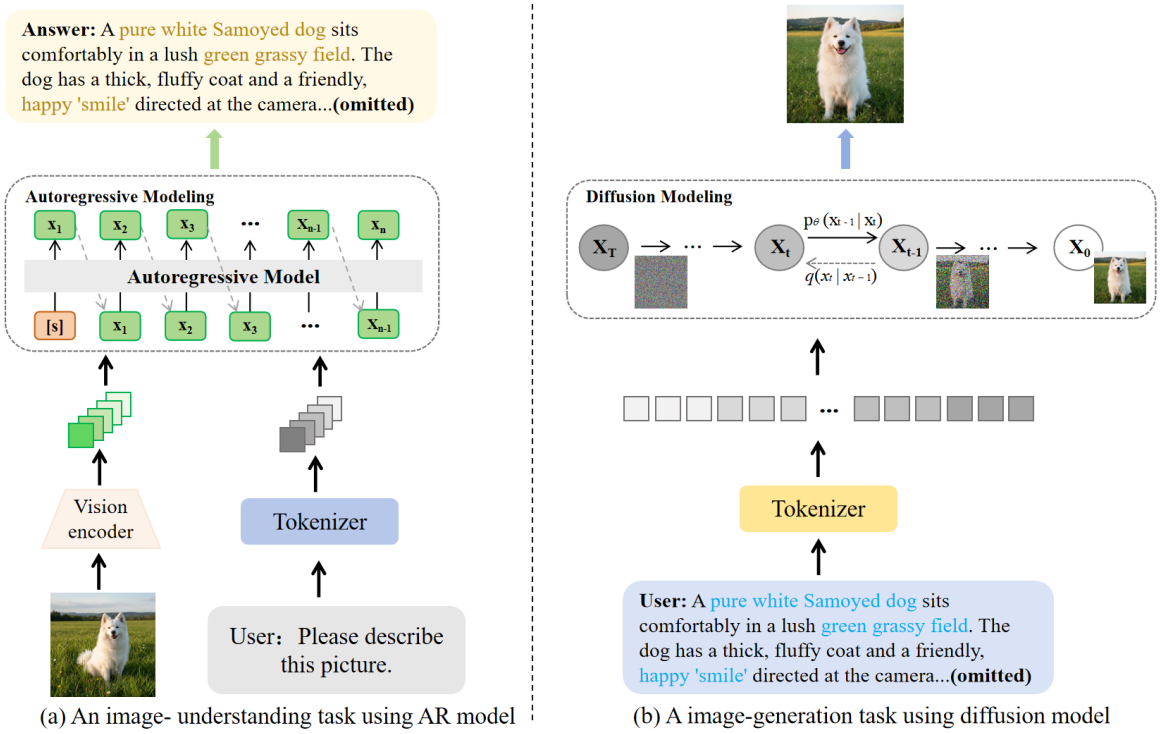

Figure 2: The illustration of Autoregressive-based Semantic Understanding and Diffusion-based Visual Generation.(a) Autoregressive (AR) modeling for image understanding, where visual features extracted by a vision encoder are aligned with textual tokens and decoded sequentially to generate semantic descriptions. (b) Diffusion-based modeling for image generation, where textual tokens produced by a tokenizer condition an iterative denoising process to synthesize visual content from noise.

residual networks to facilitate training of very deep architectures. DenseNet Huang et al. (2017) connects every layer to each other, improving compactness and reducing overfitting. A significant advancement is the CLIP model Radford et al. (2021), which uses contrastive learning to align visual and textual representations, enabling zero-shot classification, image captioning, and visual question answering. Additionally, the Segment Anything Model (SAM) Kirillov et al. (2023) has revolutionized segmentation by providing high-quality, pixel-level segmentation masks across diverse visual inputs, making it highly effective for tasks requiring precise localization.

Self-Supervised Learning-Based Methods. Unsupervised and self-supervised learning methods learn visual representations without relying on labeled data, instead exploiting the inherent structure of the data. Self-supervised learning (SSL) has gained popularity for learning robust features without annotations. The DINO model Caron et al.; Oquab et al. (2023); Siméoni et al. (2025) contrasts augmentations of the same image to generate consistent, transferable representations. Similarly, VAE (Variational Autoencoders) Kingma & Welling (2013) learns compact latent variables to model the underlying structure of images, excelling in generative tasks. These methods, such as DINO and VAE, have proven highly effective in producing generalizable representations for tasks like object detection and recognition, demonstrating the power of unsupervised learning even without large labeled datasets.

2.3 Visual Understanding Methods

Multimodal understanding models extend LLM reasoning to visual information, enabling tasks like VQA and multimodal dialogue. The core challenge is bridging continuous visual signals with the LLM's discrete token space. The first technical route is **visual encoding**. While early methods used pre-trained encoders (e.g.,

ViT Dosovitskiy (2020)) with simple projection layers Zhu et al. (2023a), more sophisticated approaches use querying transformers (e.g., Q-Former Li et al. (2023b)) or gated cross-attention Alayrac et al. (2022) to selectively distill relevant visual features into soft tokens compatible with the LLM.

The second route is **information fusion**. Initial dual-encoder architectures Lu et al. (2019); Radford et al. (2021) focused on aligning latent spaces. The modern, more effective approach is LLM-centric: visual tokens are fed directly into the LLM backbone, allowing for deep cross-modal reasoning via the LLM's internal self-attention mechanisms. This is often initialized by multi-stage alignment training Chen et al. (2024) and can be enhanced by advanced architectures like Mixture-of-Experts (MoE) Wu et al. (2024b) to improve semantic coherence.

2.4 Visual Generation Methods

Visual generation models, critical for unified systems, synthesize visual content from text. This field is dominated by two paradigms: diffusion models (DMs) and autoregressive (AR) models. DMs excel in quality and diversity, while AR models offer strong sequential consistency.

DMs model generation as an iterative denoising process Ho et al. (2020). The pivotal shift was Latent Diffusion Models (LDMs) Rombach et al. (2022), which operate in a pre-trained VAE's compact latent space, enhancing efficiency. More recently, Diffusion Transformers (DiT) Peebles & Xie (2023) treat image patches as sequences. Textual conditioning is incorporated via encoders like CLIP Radford et al. (2021) or T5 Raffel et al. (2020), and increasingly, LLMs are used for richer, more aligned conditioning Zhang et al. (2024); Wang et al. (2025a).

AR models frame image generation as a sequential prediction task Van Den Oord et al. (2016). The core challenge is discretizing visual data. While early models predicted pixels (e.g., PixelCNN Van den Oord et al. (2016)), the breakthrough came from token-based models using vector quantization (e.g., VQGAN Esser et al. (2021)). A Transformer decoder then predicts these visual tokens sequentially. Recent work improves efficiency via multi-token prediction Pang et al. (2024), coarse-to-fine strategies (e.g., VAR Tian et al. (2024)), or control mechanisms (e.g., ControlAR Li et al. (2024f)) for precise editing.

3 Unified models with continuous vision tokens

This chapter provides a comprehensive overview of **continuous tokenizers** and their corresponding **generation mechanisms**, focusing on continuous diffusion and autoregressive (AR) frameworks. We begin by introducing the theoretical foundation of continuous tokenization, followed by detailed diffusion and autoregressive formulations. Subsequent sections discuss how these mechanisms are integrated within unified multimodal architectures through serial and parallel coupling paradigms.

3.1 Preliminaries

3.1.1 Continuous Tokenizers: Foundations and Formulation

A **continuous tokenizer** bridges high-dimensional visual inputs and compact, semantically meaningful latent representations that generative models can directly operate on. Unlike discrete quantization methods, continuous tokenizers learn smooth latent manifolds, where each visual feature is represented as a differentiable vector. This continuous design preserves gradient flow and fine-grained details, making it suitable for end-to-end training in diffusion and autoregressive systems.

The *Variational Autoencoder (VAE)* provides the theoretical foundation for continuous tokenization. Given an input image x, the encoder $q_\phi(z|x)$ maps it to a latent variable z, while the decoder $p_\theta(x|z)$ reconstructs x. Training maximizes the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \,||\, p(z)), \tag{1}$$

where $p(z)$ is typically a standard Gaussian prior. Compared to discrete tokenizers such as VQ-VAE, VAEs maintain continuous, differentiable representations, allowing gradient propagation through the latent space.

3.1.2 Diffusion Generation

Diffusion models define a forward noising process and a learned reverse denoising process. Given a clean sample x_0, Gaussian noise is progressively added:

$$q(x_t|x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I\big),$$

which can be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. A neural predictor $\epsilon_\theta(x_t, t)$ learns to reverse this process by minimizing the simplified denoising loss Ho et al. (2020):

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

Latent Diffusion Models (LDM) Rombach et al. (2022) improve efficiency by operating in a learned latent space instead of pixel space. An encoder E compresses x into $z = E(x)$, diffusion is applied on z, and a decoder D reconstructs $x_0 \approx D(z_0)$. Conditioning signals (e.g., text or depth) can be injected via cross-attention, leading to controllable high-fidelity generation at reduced cost. Extensions such as FLUX further enhance training efficiency through rectified flow objectives and multi-scale latent processing.

3.1.3 Autoregressive Continuous Generation

Traditional autoregressive (AR) models factorize discrete sequences:

$$p(s_{1:N}) = \prod_{t=1}^{N} p(s_t \mid s_{<t}),$$

but for continuous latent embeddings $z_t \in \mathbb{R}^d$, direct regression may fail to capture multimodal distributions. **Continuous AR** approaches retain AR dependencies while replacing discrete categorical modeling with diffusion-based conditional generation.

Specifically, for each token z_t, we define a noisy version:

$$z_t^{(\tau)} = \sqrt{\bar{\alpha}_\tau}\, z_t + \sqrt{1-\bar{\alpha}_\tau}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

and train a denoiser conditioned on previous tokens $z_{<t}$:

$$\mathcal{L}_{\text{token}}(\theta) = \mathbb{E}_{z_t,\epsilon,\tau}\Big[\|\epsilon - \epsilon_\theta(z_t^{(\tau)}, \tau; \text{context} = z_{<t})\|^2\Big].$$

This formulation combines diffusion robustness with AR's sequential modeling power, as demonstrated in MAR Li et al. (2024e), Fluid Fan et al. (2024), and Emu Sun et al. (2023b).

3.2 Coupling Strategies for Continuous Unified Models

The integration of large language models (LLMs) with continuous visual generators defines two main paradigms: **serial coupling** and **parallel coupling**(As shown in Figure 3). They differ by how reasoning and generation interact—either sequentially or concurrently—forming a spectrum from modular to deeply unified multimodal architectures.

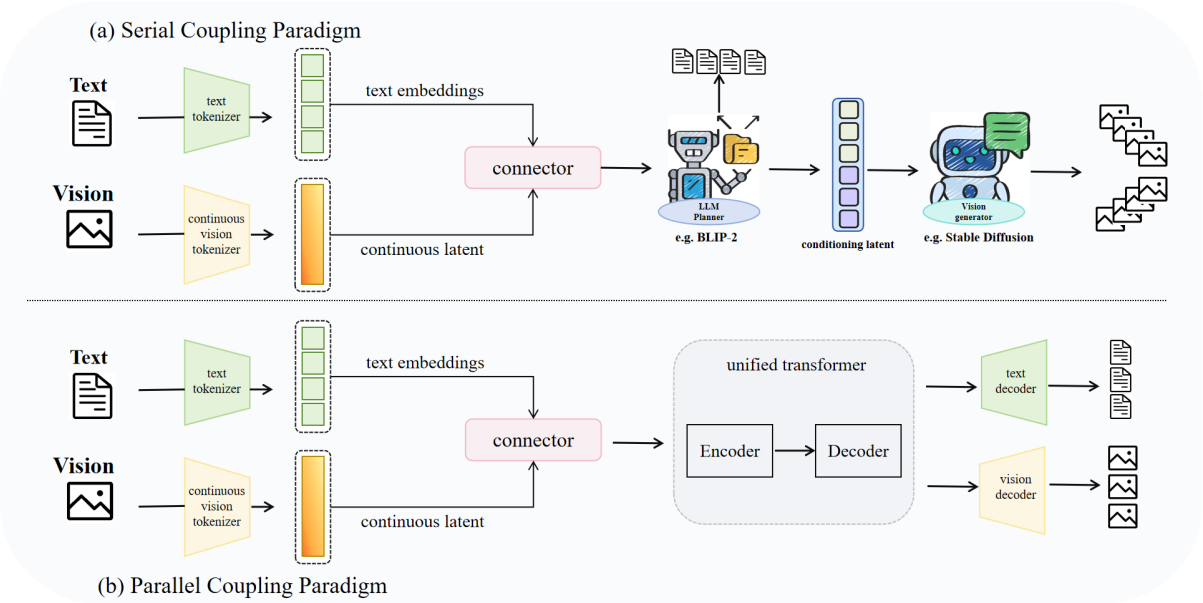

Figure 3: The illustration of serial coupling and parallel coupling. (a) Serial coupling, where multimodal inputs are first processed by a reasoning-oriented language model to produce intermediate semantic representations, which are subsequently used to condition a separate visual generation module. (b) Parallel coupling, where multimodal inputs are jointly processed within a unified transformer architecture, enabling simultaneous reasoning and visual generation through shared representations and coordinated decoding.

3.2.1 Serial Coupling Paradigm

In the **serial coupling** paradigm, the large language model (LLM) functions as a *semantic planner* that transforms high-level textual descriptions into structured conditioning signals for continuous visual generators. The generation pipeline follows a strictly modular and sequential formulation: the LLM first performs semantic reasoning and produces a latent condition c, after which a diffusion or flow-based generator consumes this representation to synthesize image latents. This design inherits the strengths of modular multimodal systems, enabling each component—textual reasoning and visual synthesis—to be independently optimized while leveraging powerful pretrained models across modalities.

In contemporary implementations, the LLM interprets and restructures user instructions into dense semantic embeddings, which are injected into latent diffusion models through cross-attention, adapter routing, or learnable query connectors. Systems such as BLIP-Diffusion Li et al. (2023b) and Emu Sun et al. (2023b) exemplify this approach: BLIP-Diffusion uses BLIP-2 as a multimodal planner to extract semantic visual features, while Emu introduces a hybrid autoregressive–diffusion mechanism where the LLM predicts coarse latent representations that are subsequently refined by a diffusion model. The SEED family Ge et al.; 2023; 2024) extends this paradigm to industrial-scale pipelines by coupling reasoning from Vicuna or LLaMA with latent diffusion backbones such as SDXL or SANA, mediated by lightweight connectors that preserve scalability. Large commercial systems—including Kosmos-G Pan et al. (2023), Nexus-Gen Zhang et al. (2025b), GPT-4o, and Gemini-1.5—also follow serial pipelines that map textual reasoning into structured latent conditions for continuous generation.

Despite its practicality and modularity, serial coupling introduces an inherent *semantic–visual gap*. Because the generation module does not provide gradient-level feedback to the LLM, mismatches may arise between the semantic planner and the visual decoder, particularly for complex layouts or fine-grained attributes. Moreover, the strictly unidirectional pipeline prevents dynamic interplay between reasoning and generation during the denoising process, limiting cross-modal coherence. While serial coupling remains favored for its

flexibility, scalability, and ease of integration with pretrained LLMs, it stops short of achieving fully unified multimodal reasoning–generation within a single model architecture.

Serial coupling ensures modularity, interpretability, and scalability, but lacks gradient feedback from generated images, creating a semantic–visual gap that limits alignment precision.

3.2.2 Parallel Coupling Paradigm

The **parallel coupling** paradigm differs fundamentally from the sequential nature of serial coupling by enabling *synchronous, bidirectional interaction* between linguistic reasoning and continuous visual generation. Instead of treating the LLM as an offline planner, parallel coupling integrates visual denoising and text reasoning within a shared or partially shared transformer architecture, allowing the two modalities to co-evolve during generation. In this formulation, text tokens typically retain causal attention to preserve autoregressive language modeling, while visual latent tokens employ bidirectional attention to accommodate diffusion or flow dynamics. The shared attention space ensures that linguistic semantics and visual structure continuously influence each other, thereby achieving deeper multimodal unification.

This paradigm is exemplified by models such as Transfusion Zhou et al. (2024) and MonoFormer Zhao et al. (2024), which adopt unified transformer backbones with distinct attention masks for language and vision. In Transfusion, autoregressive text generation and latent diffusion denoising are fused within a single architecture: language tokens are processed with causal masking whereas image tokens use full bidirectional attention, with cross-attention pathways enabling mutual guidance. MonoFormer follows a similar design but simplifies the multimodal integration through shared layers and structured attention routing, providing efficient interleaving of text reasoning and latent-space image generation. These models demonstrate how integrating text and vision streams at every generation step provides stronger semantic alignment than a serial pipeline could achieve.

LMFusion Shi et al. (2024) extends the parallel paradigm by adapting pretrained LLMs to serve as the backbone for multimodal interaction. Lightweight visual modules and shared attention layers are introduced to merge visual features into the frozen language backbone, enabling the model to leverage the stability of LLM parameters while achieving effective bidirectional fusion between reasoning and perception. BAGEL Deng et al. (2025) generalizes this idea via a Mixture-of-Transformers (MoT) framework. It separates modality- and task-specific attention routes while maintaining a unified architecture, allowing flexible yet coherent interactions across tasks such as image generation, editing, captioning, and dialogue. Although powerful, these architectures require substantial computation due to the need for cross-modal synchronization and large unified transformer backbones.

A notable recent advancement within the parallel coupling landscape is Skywork- UniPic Wang et al. (2025b), which demonstrates that unified autoregressive modeling can support visual understanding, generation, and editing within a single architecture. UniPic employs a decoupled but jointly trained visual encoding pathway—a Masked Autoregressive (MAR) encoder-decoder for generation and a SigLIP2 encoder for perception—while sharing a single language backbone across tasks. This design allows text reasoning and visual latent evolution to take place concurrently, embodying the essence of parallel coupling: language tokens guide visual synthesis at each step, while image latent tokens reciprocally inform linguistic processing. UniPic highlights that continuous visual generation and language modeling can be co-trained and co-executed without handcrafted connectors, reinforcing the shift toward tightly unified multimodal systems.

Overall, parallel coupling offers significantly enhanced cross-modal coherence, enabling functionalities such as real-time editing, interactive visual dialogue, instruction-driven fine-grained manipulation, and open-ended multimodal reasoning. However, such tightly integrated architectures come with increased training and inference costs, complex token routing requirements, and higher sensitivity to optimization dynamics. Despite these challenges, parallel coupling represents a critical step toward fully unified multimodal foundation models capable of seamless joint reasoning and generation.

Table 1: Comparison between the **Serial Coupling** and **Parallel Coupling** paradigms for unified continuous visual models.

Aspect	Serial Coupling	Parallel Coupling
Core Philosophy	LLM acts as a *semantic planner*; reasoning and generation occur sequentially.	Text reasoning and visual denoising occur *synchronously* within shared or partially shared modules.
Information Flow	Unidirectional: LLM → visual generator.	Bidirectional: language and vision tokens interact at every step of generation.
Representation Interaction	Loose coupling; cross-attention or adapters provide semantic constraints.	Tight coupling; shared transformer blocks or unified attention layers align modalities.
Generative Process	Visual generation is conditioned on static semantic embeddings.	Visual denoising dynamically conditions on evolving language states, and vice versa.
Advantages	Modular, scalable, easy to integrate with pretrained LLMs; efficient for large-scale deployment.	Strong cross-modal coherence; supports interactive editing, multimodal reasoning, and fine-grained control.
Limitations	Semantic–visual gap; no gradient feedback from image to LLM; limited real-time alignment.	High computational cost; complex attention scheduling; difficult optimization due to tight coupling.
Representative Systems	BLIP-Diffusion Li et al. (2023b), Emu Sun et al. (2023b), SEED Ge et al.; 2023; 2024), Kosmos-G Pan et al. (2023), Nexus-Gen Zhang et al. (2025b).	Transfusion Zhou et al. (2024), Mono-Former Zhao et al. (2024), LMFusion Shi et al. (2024), BAGEL Deng et al. (2025), UniPic Wang et al. (2025b).
Suitable Use Cases	High-throughput generation, instruction-to-image pipelines, modular system design.	Unified multimodal modeling, real-time multimodal tasks, understanding–generation joint optimization.

3.2.3 Analysis and Outlook

Serial coupling prioritizes scalability and modularity, while parallel coupling maximizes coherence through shared representation learning. Emerging hybrid models combine both—serial pipelines with partial parallel feedback—to balance efficiency and multimodal alignment. Future directions point toward *adaptive coupling*, dynamically controlling the degree of inter-modality interaction based on task context, progressing toward fully unified multimodal foundation models.

3.3 Challenges for Continuous Unified Models

Continuous visual representations bring advantages such as high perceptual fidelity, differentiability, and natural compatibility with diffusion and flow models. Nonetheless, they also present challenges: weak interpretability, substantial computational cost, and non-trivial semantic alignment between continuous vision and discrete language spaces. Addressing these limitations remains central to advancing continuous unified multimodal systems.

4 Unified models with discrete vision tokens

Unified multimodal models employing discrete visual generation represent a powerful paradigm where both understanding and generation operate within a shared, tokenized conceptual space. Unlike continuous generation methods that manipulate latent vectors, these models treat visual content—much like text—as sequences of discrete tokens. This approach naturally aligns with the sequence-to-sequence nature of Large Language Models (LLMs), allowing the entire understanding-to-generation pipeline to be executed by a single, often Transformer-based, backbone.

The primary motivation for adopting discrete visual representations lies in their ability to leverage the powerful capabilities of LLMs, which have been extensively pre-trained on vast amounts of textual data. By transforming visual content into discrete tokens, these models enable the integration of vision tasks with the well-established techniques used for language tasks, facilitating the unification of understanding and generation. This tokenization simplifies the bridging of the semantic gap between language and vision, as it allows the use of a shared framework that processes both modalities in a similar manner.

Furthermore, discrete visual representations bring significant advantages in terms of **efficiency** and **coherence** with language. By converting visual content into a set of discrete tokens, visual information is structured in a way that aligns more directly with the discrete nature of text. This structural alignment allows vision tasks to be integrated seamlessly into a unified framework, where both visual and language models operate in a similar tokenized space. As a result, discrete models facilitate **cross-modal coherence**, making it easier for the system to handle both vision and language tasks simultaneously, with the visual content serving as tokens that can be processed in parallel with text. This approach enables **modularization** of the vision and language components. Each modality can be independently optimized and improved, while still ensuring that they can interact effectively within the same architecture. Additionally, discrete representations are often more **computationally efficient** compared to continuous models. The discrete tokens are easier to process and store, which reduces the computational burden, especially in large-scale multimodal tasks. By simplifying the representation of visual data, discrete models enable efficient scaling to high-dimensional visual and textual data without compromising the **alignment** between the two modalities.

This chapter first explores the critical process of converting continuous visual signals into discrete representations. Subsequently, it delves into the two primary methodologies for generating these discrete visual tokens within unified frameworks(As show in figure 4): autoregressive-driven generation and discrete diffusion-driven generation.

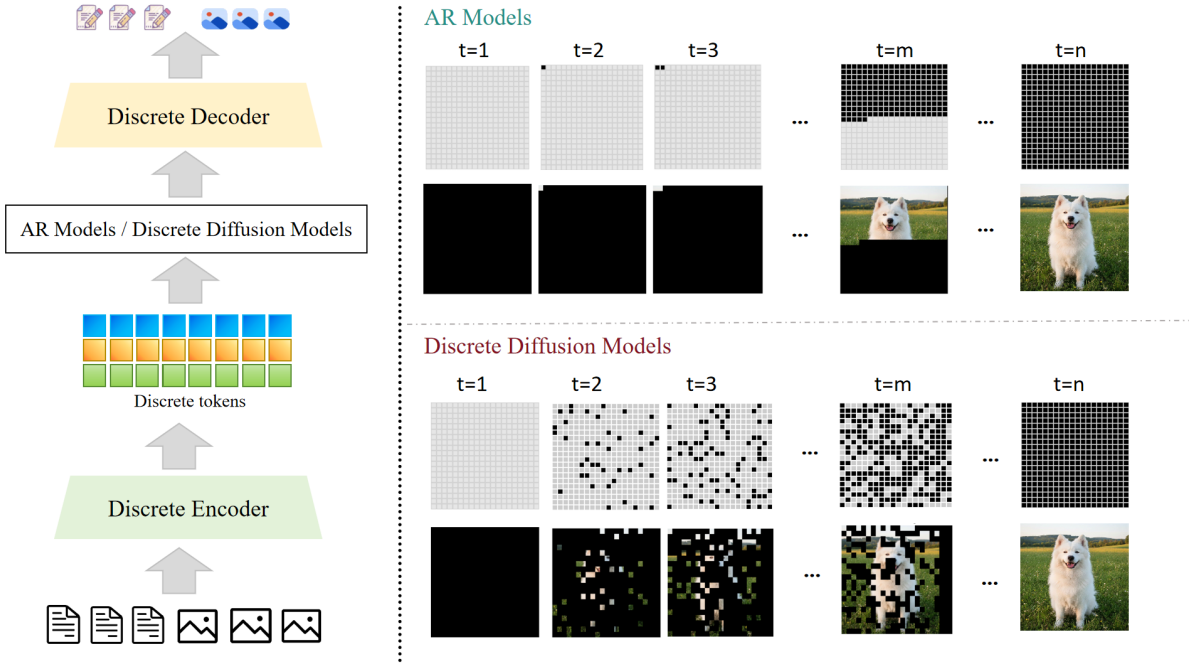

Figure 4: Comparison of Autoregressive and Diffusion Paradigms for unified models with discrete vision tokens.

4.1 Visual Vector Quantization

A unified multimodal framework hinges on the ability to represent continuous visual signals—such as images, videos, or spatial-temporal scenes—as discrete sequences of tokens. This transformation enables vision data

to share the same representational format as language, allowing both understanding and generation to be handled by a single sequence model. The challenge lies in designing discrete representations that are compact, expressive, and semantically aligned across modalities.

4.1.1 Preliminary on Vector Quantization

Most discrete visual representations stem from the principle of vector quantization (VQ), first introduced by *VQ-VAE* Van Den Oord et al. (2017). In these frameworks, an encoder $\text{Enc}(\cdot)$ maps the input x into a latent feature map $z = \text{Enc}(x)$, which is then discretized by replacing each latent vector with its nearest entry from a learned codebook $\mathcal{C} = c_1, \ldots, c_M$:

$$Q(z_i) = \arg\min_{c_j \in \mathcal{C}} \|z_i - c_j\|_2.$$

A decoder $\text{Dec}(\cdot)$ reconstructs $\hat{x} = \text{Dec}(Q(z))$, and the entire system is trained to minimize reconstruction and commitment losses. This process produces discrete code indices that serve as "visual tokens," analogous to word tokens in NLP.

The key challenge in VQ is managing the trade-off between compression and fidelity. *VQ-GAN* Esser et al. (2021), for instance, incorporates adversarial and perceptual losses to recover high-frequency details. Subsequent refinements have focused on common failure modes: hierarchical latents (e.g., VQ-VAE-2 Razavi et al. (2019)) capture multi-scale features, while various regularization techniques (e.g., Reg-VQ Zhang et al. (2023a), HC-VQ Volkov (2022)) and re-initialization strategies (e.g., HQA Williams et al. (2020), CVQ-VAE Zheng & Vedaldi (2023)) are used to prevent codebook collapse. Soft-assignment methods (e.g., SQ-VAE Takida et al. (2022), SCQ Gautam et al. (2023)) have also been explored to create richer, less-constrained representations.

4.1.2 Advanced Variants of Vector Quantization

Residual Vector Quantization (RVQ) decomposes latent vectors into multiple residual stages, progressively refining quantized codes by quantizing the residual error from previous stages. This hierarchical approach allows for multi-scale representation, capturing both coarse and fine details Barnes et al. (1996); Chen et al. (2010). Variants primarily focus on optimizing the stages, such as using hierarchical dependency (SQ Martinez et al. (2014)), joint optimization (ErVQ Ai et al. (2017)), or subspace clustering (IRVQ Liu et al. (2015)). Recent approaches like QINCo Huijben et al. (2024) use neural networks to generate the step-specific codebooks.

Product Quantization (PQ) partitions the latent space into independent subspaces, each with its own sub-codebook, greatly expanding the effective vocabulary size without linear parameter growth. This method is particularly efficient for high-dimensional visual data Jegou et al. (2010); Matsui et al. (2018). Extensions focus on optimizing the subspace decomposition (OPQ Ge et al. (2013), LOPQ Kalantidis & Avrithis (2014)), enabling online updates (Online PQ Xu et al. (2018)), or making the process differentiable (DPQ Klein & Wolf (2019), DOPQ Lu et al. (2023b)).

Additive Vector Quantization (AQ) represents vectors as sums of codewords to promote sparsity and efficiency Babenko & Lempitsky (2014). Variants include LSQ Martinez et al. (2016), which uses iterated local search for encoding. Recent developments, such as **Finite Scalar Quantization (FSQ)** Mentzer et al. (2023) and **Lookup-Free Quantization (LFQ)** Yu et al. (2023a), avoid explicit codebooks by using rounding or sign-based decompositions. These approaches (e.g., LFQ as used in MAGVIT-v2 Yu et al. (2023a)) mitigate codebook collapse and improve training stability, facilitating end-to-end optimization.

4.1.3 Challenges and Future Directions

Challenges. One primary challenge is the *codebook collapse* problem, where certain visual tokens are never utilized during training, leading to inefficient representations. This is particularly problematic when attempting to align visual and language tokens in a unified space. Furthermore, the *quantization error* introduced during the conversion of continuous visual signals to discrete tokens can accumulate, affecting both understanding and generation tasks. This issue is especially prominent in tasks that require high fidelity,

such as image generation or fine-grained reasoning. Another challenge is the *alignment between modalities*, where token granularity may need to be dynamically adjusted depending on the task—fine-grained tokens for generation and coarser tokens for understanding. The lack of *semantic alignment* across modalities can hinder multimodal integration, particularly when visual features do not correspond clearly to textual representations.

Future Directions. Future work should focus on *adaptive quantization mechanisms* that adjust token granularity based on the specific task or modality. Additionally, developing *unified token spaces* where both vision and language tokens can co-exist and interact seamlessly will be crucial for improving multimodal models. Incorporating *hybrid encoding schemes* that combine discrete tokens with continuous features will help maintain both high-level semantics and low-level details. Lastly, efforts should be directed towards improving *model controllability* and *explainability*, enabling more precise generation and better interpretability of the model's decisions.

4.2 Autoregressive-Driven Vision-Language Unifying

The narrative of autoregressive (AR) models in unified multimodal systems unfolds as a compelling extension of language modeling's triumphs into the visual realm. Originating from the sequential prediction paradigms that powered early successes in natural language processing—such as GPT's next-word forecasting—AR approaches adapted this causality to visuals, treating images and videos as token streams ripe for step-by-step construction. This evolution, as chronicled in recent surveys Zhang et al. (2025c); Li et al. (2025a), marks a pivotal shift: from isolated visual tasks to a harmonious unification where understanding (e.g., scene reasoning) emerges as a byproduct of generative prediction. By modeling visuals autoregressively, these models not only generate coherent outputs but also infer semantics through the very act of forecasting, bridging the discrete worlds of pixels and words.

The story begins with foundational challenges: how to serialize high-dimensional visuals without losing fidelity, and how to fuse them with language for true multimodality. Early AR models tackled pixel-level generation, but as scales grew, innovations in tokenization and integration propelled unification forward. This section traces this arc through a structured classification, drawing on Zhang et al. (2025c)'s emphasis on AR's role in large vision models (LVMs) and Li et al. (2025a)'s insights into discrete generative paradigms. We categorize along three dimensions—representation strategies (encoding strategies), modality fusion mechanisms, and training methodologies—highlighting key milestones that transformed AR from a generative tool into a unified powerhouse.

4.2.1 Fundamental Mechanism

Autoregressive visual generators model the joint distribution of a tokenized image $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ as

$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid x_{<i}, c),$$

where c denotes conditional inputs such as textual prompts, prior visual context, or multimodal embeddings. After generating a complete token sequence, a learned decoder (e.g., from a VQ-VAE or VQ-GAN) reconstructs the final image from discrete code indices. This causal formulation directly parallels the decoding process of large language models (LLMs), enabling a unified modeling interface for both textual and visual outputs. In recent unified models, AR modeling achieves unification by treating understanding and generation as variants of next-token prediction, naturally bridging discrete visual tokens with linguistic sequences.

4.2.2 Categories of Unified Autoregressive Models

Autoregressive (AR) unified models differ primarily in how they represent and serialize visual information before integration with language tokens. While Chapter **??** introduced the fundamental encoding paradigms—pixel-based, semantic-based, and hybrid tokenization—this section categorizes representative

unified AR models according to the encoding strategy they adopt and how it defines their overall modeling behavior. Each category reflects a distinct trade-off among fidelity, alignment, and efficiency.

(1) Pixel-Token Unified Models. These models treat visual content as sequences of low-level discrete tokens obtained from pixel-oriented quantizers such as VQ-VAE or VQ-GAN. The tokens are serialized in raster order and processed autoregressively together with text tokens, forming a fully unified sequence-to-sequence pipeline. Representative works include Chameleon Team (2024), ANOLE Chern et al. (2024), Emu3 Wang et al. (2024), UGen Tang et al. (2025), and Liquid Wu et al. (2024a). This design enables fine-grained visual reconstruction and photorealistic synthesis within a single Transformer backbone. However, pixel-token unification typically leads to extremely long token sequences—growing quadratically with image resolution—and the tokens themselves carry limited semantic abstraction. Consequently, while these models achieve high visual fidelity, they often struggle with cross-modal reasoning, long-context understanding, and computational scalability.

(2) Semantic-Token Unified Models. A second line of work encodes images using text-aligned encoders such as CLIP, SigLIP, or EVA-CLIP, transforming them into semantically meaningful discrete embeddings. In this formulation, both visual and textual tokens inhabit a shared conceptual space, and the AR model operates directly on semantically aligned sequences. Representative systems include LaViT Jin et al. (2023), DreamLLM Dong et al. (2023), VL-GPT Zhu et al. (2023b), and MetaMorph Tong et al. (2025). This paradigm excels in cross-modal understanding and instruction following, as semantic embeddings align naturally with linguistic structure. However, the abstraction that enables strong reasoning also sacrifices spatial precision and pixel-level controllability, requiring separate diffusion or reconstruction decoders for high-quality generation. Thus, semantic-token unification emphasizes reasoning and understanding rather than direct synthesis.

(3) Hybrid-Token Unified Models. Hybrid models aim to combine the complementary strengths of pixel-level fidelity and semantic-level abstraction by fusing both types of tokens. Two primary variants exist: (a) *pseudo-hybrid* systems that employ distinct encoders for different tasks (e.g., semantic tokens for understanding, pixel tokens for generation), and (b) *joint-hybrid* systems that deeply integrate both token types into a single sequence space. Examples include Janus and Janus-Pro Wu et al. (2025); Chen et al. (2025b), OmniMamba Zou et al. (2025), MUSE-VL, SemHiTok Chen et al. (2025c), and Show-o2 Xie et al. (2025). These approaches achieve more balanced performance between understanding and generation and demonstrate improved adaptability across diverse multimodal tasks. Nonetheless, they introduce new challenges in feature fusion, token redundancy, and training efficiency, as maintaining coherence across multiple token spaces remains non-trivial.

(4) Summary and Discussion. Autoregressive unified models have demonstrated the feasibility of handling visual and linguistic sequences within a single generative backbone. Yet, their architectural nature brings several inherent limitations:

- **Sequential inefficiency:** Token-by-token decoding causes inference latency to scale linearly with sequence length, which becomes prohibitive for high-resolution visual data.
- **Local generation without global refinement:** The strictly causal generation process prevents global feedback or holistic correction, leading to artifacts or incoherent layouts.
- **Semantic–structural imbalance:** Pixel-token models capture fine details but lack conceptual grounding, while semantic-token models achieve reasoning at the expense of controllable visual precision.

These issues motivate the exploration of *non-autoregressive paradigms*—particularly discrete diffusion models—that can perform **parallel, globally coherent, and semantically consistent** generation in a shared token space. The next section introduces this emerging direction, where discrete diffusion serves as a more flexible and efficient mechanism for unified vision–language modeling.

4.3 Discrete Diffusion-Driven Vision–Language Unifying

Foundations. In contrast to the autoregressive (AR) paradigm described in Section 4.2.2, diffusion-based modeling offers a fundamentally different view of generation. Instead of predicting the next token causally, diffusion reformulates generation as an *iterative denoising process*—a trajectory from structured noise to clean data. This paradigm is inherently non-sequential and global, enabling parallel decoding and holistic consistency.

Early works on continuous diffusion, such as DDPM and Latent Diffusion Models, have dominated pixel-level generation. However, extending diffusion to discrete token spaces—essential for unified vision–language modeling—required a series of conceptual breakthroughs. The seminal D3PM (Discrete Denoising Diffusion Probabilistic Model) Austin et al. (2021) first introduced a categorical noise formulation:

$$q(x_t \mid x_{t-1}) = (1 - \beta_t)\,\mathrm{Cat}(x_{t-1}) + \beta_t\, p_{\text{noise}}(x),$$

where β_t controls the corruption rate and $p_{\text{noise}}(x)$ defines the discrete noise prior (e.g., uniform or absorbing-state). This formulation preserves the Markovian structure of continuous diffusion while supporting discrete vocabularies such as text or visual tokens.

Subsequent work, SEDD (Score Entropy Discrete Diffusion) Lou et al. (2023), refined this framework by introducing entropy-based training objectives for better stability and sample quality, establishing diffusion as a viable alternative to AR decoding in discrete domains. These advances bridged the conceptual gap between continuous image diffusion and token-based generative modeling.

Building upon these foundations, LLaDA (Large Language and Discrete Diffusion Alignment) Nie et al. (2025) demonstrated that discrete diffusion can be scaled to LLM-level architectures. By treating both textual and visual tokens as elements of a shared categorical vocabulary, LLaDA unified multimodal denoising under a single Transformer backbone, revealing the potential of diffusion as a *non-causal, globally coherent unifier* for multimodal generation.

4.3.1 Fundamental Mechanism.

In unified discrete diffusion, the forward process defines a discrete Markov chain $\{x_t\}_{t=0}^{T}$ as above, and the reverse process—parameterized by a Transformer—predicts $p_\theta(x_{t-1} \mid x_t)$ at each step. Two corruption schemes are dominant:

- **Absorbing-state diffusion:** Tokens are replaced with a special absorbing symbol such as `[MASK]` with probability β_t. Once absorbed, a token remains masked until recovered during denoising. This design yields a well-defined termination state and smooth training dynamics.
- **Uniform-state diffusion:** Tokens are randomly replaced with uniformly sampled vocabulary items, producing a fully mixed state as $t \to T$. This variant simplifies the forward process but requires stronger modeling capacity in the reverse direction.

The model is trained via cross-entropy between predicted and ground-truth clean tokens, which serves as a discrete analogue of score matching.

From Discrete Diffusion to Unified Multimodality. The key insight enabling unified vision–language diffusion is that both visual and textual modalities can be represented as discrete tokens—e.g., from VQ-VAE or semantic tokenizers—thus sharing the same categorical domain. This allows a single diffusion Transformer to jointly denoise across modalities under *full attention*, enabling bidirectional information flow between text and vision. In contrast to AR models, diffusion updates all tokens in parallel at each timestep, facilitating:

- **Parallel decoding** — multi-token updates significantly reduce inference latency, especially for long visual sequences.
- **Full-context attention** — bidirectional attention allows global semantic reasoning and cross-modal coherence.

- **Iterative global refinement** — each denoising step refines prior outputs, ensuring consistent spatial and semantic alignment.

4.3.2 Evolution of Unified Diffusion Models.

Following the theoretical establishment of discrete diffusion, a new generation of models extended these ideas into large-scale unified vision–language systems: UniDisc Swerdlow et al. (2025) pioneered the notion of unified discrete diffusion, introducing a masking-based denoising objective that jointly handles text and image tokens within one diffusion framework. It demonstrated controllable, non-autoregressive generation and editing, proving the feasibility of full unification. MMaDA Yang et al. (2025b) advanced this paradigm toward reasoning-capable multimodal diffusion. By integrating chain-of-thought (CoT) reasoning and textual guidance during denoising, MMaDA bridged structured cognition and diffusion-based generation. Lumina-DiMOO Xin et al. (2025) further scaled unified diffusion to the *"omni-modeling"* level, leveraging fully discrete tokenization and efficient sampling schemes to achieve high fidelity and sampling speed. Muddit Shi et al. (2025) focused on improving visual realism by injecting priors from pre-trained text-to-image diffusion backbones, thus enhancing fidelity while maintaining unified control. Lavida-O Li et al. (2025b) introduced an Elastic-MoT (Mixture-of-Tasks) architecture that supports high-resolution (1024px) synthesis, object-level grounding, and compositional reasoning—signaling the maturation of unified discrete diffusion into scalable multimodal intelligence.

4.3.3 Summary and Outlook.

The progression from foundational discrete diffusion models (D3PM, SEDD) to large-scale unified frameworks (LLaDA, UniDisc, MMaDA, Lumina-DiMOO, Lavida-O) marks a paradigm shift from sequential to globally consistent multimodal generation. By reformulating autoregressive prediction as iterative denoising, diffusion-driven unification overcomes key AR limitations—achieving *parallel inference*, *bidirectional semantic alignment*, and *iterative refinement* within a single generative process. This paradigm not only provides theoretical elegance but also practical flexibility, supporting diverse tasks such as image synthesis, captioning, editing, and multimodal reasoning.

Despite these advances, the community surrounding discrete diffusion models is still evolving, and several challenges remain. Notably, the lack of a mature ecosystem for model training and experimentation means that getting robust results from scratch can be challenging. Training discrete diffusion models from zero is often less effective compared to pre-trained models, as the learning process can be highly sensitive to data quality, model architecture, and hyperparameters. Additionally, scaling models to handle more complex multimodal tasks can expose issues related to optimization stability and computational efficiency. These hurdles make it clear that while discrete diffusion is emerging as a powerful tool for holistic, controllable, and scalable vision-language intelligence, further research and community collaboration are needed to overcome the current limitations and refine these systems for broader use.

4.4 Challenges of Discrete Unified Models

Discrete visual unified models, while aligning well with LLM architectures, face trade-offs in generative quality. The use of discrete representations often leads to a loss of fine visual details due to quantization, resulting in lower fidelity compared to continuous methods. This compromises the sharpness and realism of generated images.

Additionally, these models can struggle with handling complex visual details that require high-resolution generation, making them less effective in certain tasks where fine-grained outputs are crucial. Despite these drawbacks, the alignment with LLMs remains a key advantage, enabling more seamless multimodal integration and joint reasoning across vision and language tasks.

While these models offer a promising path towards scalable, unified frameworks, further research is needed to balance the trade-off between quality and multimodal alignment.

Table 2: Concise comparison between **Autoregressive** and **Discrete Diffusion** paradigms for unified discrete vision–language models.

Aspect	Autoregressive (AR)	Discrete Diffusion
Generation	Sequential next-token prediction	Parallel iterative denoising
Attention Pattern	Causal (unidirectional)	Full (bidirectional)
Global Consistency	Limited; errors accumulate	Strong; iterative global refinement
Speed	Slow for long visual sequences	Faster via parallel updates
LLM Compatibility	Naturally aligned with LLM training	Requires diffusion-specific objectives
Visual Fidelity	High with pixel tokens, but expensive	High with global refinement, scalable
Cross-Modal Fusion	Via concatenation / causal fusion	Joint denoising across modalities
Strengths	Simple, unified interface; strong reasoning	Coherent synthesis; robust editing abilities
Limitations	Sequential bottleneck; weak global structure	Harder training; less mature ecosystem

5 Datasets and benchmarks

5.1 Datasets for Unified Multimodal Models

A unified multimodal model's success crucially depends on the diversity and scale of its training data, which must support both perception and generation. These datasets can be broadly categorized into four major groups: (1) **Multimodal Understanding** datasets (e.g., LAION, COYO) for vision–language alignment; (2) **Text-to-Image Generation** datasets (e.g., LAION-Aesthetics, Echo-4o-Image) for synthesizing high-quality visual content; (3) **Interleaved Image–Text** datasets (e.g., OBELICS, OmniCorpus) for providing document-level multimodal contexts; and (4) **Image Editing** datasets (e.g., InstructPix2Pix, AnyEdit) for controllable manipulation via instructions.

Table 3 summarizes representative datasets across these categories, which are often used in combination. Unified models typically rely on large-scale understanding corpora for grounding, generative datasets for synthesis, interleaved data for context reasoning, and editing datasets for controllable manipulation. The curation, filtering, and balancing of these multimodal datasets remain central to achieving robust unification.

Table 3: Representative Datasets for Unified Multimodal Models

Category	Dataset Example	Ref.	Primary Contribution
Understanding	LAION / COYO	Schuhmann et al. (2022); Byeon et al. (2022)	Massive-scale (billions) web-scraped image-text pairs.
	DataComp	Gadre et al. (2023)	Standardized large-scale (1.4B) data curation benchmark.
	LLaVA-OneVision	Li et al. (2024b)	Large-scale (4.8M) instruction-following data.
Generation	LAION-Aesthetics	Schuhmann et al. (2022)	Filtered subset (120M) for high-quality T2I generation.
	AnyWord-3M / CosmicMan	Tuo et al. (2023); Li et al. (2024d)	Focus on compositional reasoning and character consistency.
	Echo-4o-Image / BLIP-3o	Ye et al. (2025); Chen et al. (2025a)	Instruction-tuned data for unified generative alignment.
Interleaved	OBELICS / OmniCorpus	Laurençon et al. (2023); Li et al. (2024c)	Web-scale (141M / 8B) interleaved image-text documents.
	Multimodal C4	Zhu et al. (2023c)	Document-level multimodal contexts (101M documents).
Editing	InstructPix2Pix	Brooks et al. (2023)	Paired (image, instruction, edited_image) data.
	SEED-Data-Edit / AnyEdit	Ge et al. (2024); Yu et al. (2025)	Large-scale (3.7M+) instruction-driven edit corpora.
	ByteMorph	Chang et al. (2025)	Fine-grained non-rigid deformation editing examples.

5.2 Benchmarks for Unified Multimodal Models

The evaluation of unified models requires benchmarks that span both understanding and generation. These benchmarks collectively form a continuous evaluation spectrum from perception to synthesis and are summarized in Table 4. Key evaluation categories include: (1) **Understanding**, which tests perception and reasoning (e.g., VQA, SEED-Bench); (2) **Image Generation**, assessing quality and controllability (e.g., DrawBench, T2I-CompBench); and (3) **Interleaved Generation**, which evaluates the ability to interleave text and visual reasoning (e.g., InterleavedBench, UniBench).

Table 4: Key Benchmarks for Unified Multimodal Models

Category	Benchmark Example	Ref.	Primary Evaluation Focus
Understanding	VQA / OK-VQA	Antol et al. (2015); Marino et al. (2019)	Visual question answering (standard and knowledge-based).
	MM-Vet / SEED-Bench	Yu et al. (2023b); Li et al. (2023a)	Multi-domain, fine-grained multimodal comprehension.
	MathVista / General-Bench	Lu et al. (2023a); Fei et al. (2025)	General-purpose reasoning (visual, textual, mathematical).
Generation	DrawBench / PartiPrompts	Saharia et al. (2022); Yu et al. (2022)	Prompt fidelity and complex scene composition.
	T2I-CompBench	Huang et al. (2023)	Compositional reasoning (e.g., attributes, spatial relations).
	GenAI-Bench / WorldGenBench	Li et al. (2024a); Zhang et al. (2025a)	Standardized protocols for T2I generation quality.
Interleaved	InterleavedBench	Liu et al. (2024)	Human-curated samples of interleaved text and images.
	OpenLEAF / OpenING	An et al. (2023); Zhou et al. (2025)	Open-domain query-response and mixed-modality generation.
	UniBench	Li et al. (2025c)	Fine-grained evaluation of text-image-text compositions.
Editing / Other	MagicBrush	Zhang et al. (2023b)	Instruction-based real-image editing.
	DreamBench++	Peng et al. (2024)	Subject-driven and identity-consistent generation.
	VTBench	Lin et al. (2025)	Visual tokenizer reconstruction quality (for discrete models).

Table 5: Comparison of unified multimodal models with continuous and discrete vision tokens for zero-shot understanding across multiple datasets. **Bold** is best and underline is second best.

Model	VQAv2	OK-VQA	GQA	NoCaps	Flickr	MMMU	MMB	MME-P
continuous vision tokens								
BLIP-2 Li et al. (2023b)	65.0	45.9	44.7	**121.6**	**97.6**	-	-	-
Emu Sun et al. (2023b)	52.0	38.2	-	96.5	72.0	-	-	-
Emu-I Sun et al. (2023b)	57.2	43.4	-	108.8	77.4	-	-	-
SEED-LLaMA Li et al. (2023a)	44.2	29.2	-	-	-	-	-	-
SEED-LLaMA-I Ge et al. (2024)	66.2	45.9	-	-	-	-	-	-
Nexus-Gen 7B Zhang et al. (2025b)	**79.3**	-	-	-	-	45.7	-	1602.3
LLaVAFusion Zhou et al. (2024)	-	-	-	-	-	41.7	-	1603.7
BAGEL Deng et al. (2025)	-	-	-	-	-	43.2	79.2	1610
discrete vision tokens								
Chameleon-MultiTask Team (2024)	69.6	-	-	-	76.2	-	-	-
Emu3 Wang et al. (2024)	75.1	-	60.3	-	76.2	31.6	58.5	1243.8
Liquid Wu et al. (2024a)	68.0	-	56.1	-	-	-	-	1119.3
LaViT Jin et al. (2023)	66.0	**54.6**	46.8	114.2	83.0	-	58.0	-
Dream-LLM Dong et al. (2023)	56.6	44.3	46.8	114.2	83.0	-	-	-
VL-GPT Zhu et al. (2023b)	51.7	35.8	34.6	-	-	-	-	-
Janus Wu et al. (2025)	51.7	35.8	34.6	-	-	30.5	69.4	1338.0
SemHiTok Chen et al. (2025c)	-	-	60.3	-	-	39.3	72.3	1449.0
show-o Xie et al. (2025)	69.4	-	58.0	-	62.5	26.7	-	1097.2
show-o2 Xie et al. (2025)	-	-	**63.1**	-	-	**48.9**	**79.3**	**1620.5**
MMaDA Yang et al. (2025b)	76.7	-	61.3	-	67.6	30.2	68.5	1410.7

5.3 Comparative Analysis of unified models on Evaluation Tasks

Table 5 presents the performance of unified multimodal models with continuous and discrete vision tokens across a diverse set of multimodal understanding benchmarks, including VQAv2Antol et al. (2015), OK-VQAMarino et al. (2019), GQAHudson & Manning (2019), NoCapsAgrawal et al. (2019), Flickr30KYoung et al. (2014), MMMUYue et al. (2023), MMBLiu et al. (2023), and MME-PFu et al. (2023).The results reveal that models based on continuous vision tokens, such as BLIP-2, Emu, and Emu-I, generally show stronger capability on captioning-oriented tasks, with BLIP-2 achieving the highest CIDEr and TR@1 scores on NoCaps and Flickr respectively. Meanwhile, Nexus-Gen and BAGEL demonstrate competitive performance on broader reasoning benchmarks such as MMMU and MME-P. In contrast, discrete-token models—including LaViT, Emu3, and show-o2—excel on structured reasoning tasks such as GQA, MMMU, and MMB, with show-o2 achieving the strongest overall results across multiple metrics.

Table 6 compares representative models on the GenEval Bakhtin et al. (2018) benchmark for evaluating text-to-image generation ability. Continuous-token models such as SEED-X, Nexus-Gen, GPT-4o, and BAGEL achieve strong object-level fidelity, accurate color rendering, and stable attribute control, with BAGEL and GPT-4o delivering consistently high scores across most categories. In contrast, discrete-token models, in-

Table 6: Evaluation of text-to-image generation ability on GenEval benchmark. **Bold** is the best and underline is the second best.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
continuous vision tokens							
SEED-X Ge et al. (2024)	0.97	0.58	0.26	0.80	0.19	0.14	0.49
Nexus-Gen Zhang et al. (2025b)	0.99	0.86	0.53	0.85	0.78	0.59	0.77
GPT-4o Hurst et al. (2024)	0.99	0.92	**0.85**	**0.92**	0.75	0.61	0.84
BAGEL Deng et al. (2025)	0.99	**0.94**	0.81	0.88	0.64	0.63	0.82
Skywork unipic Wang et al. (2025b)	0.98	0.92	0.74	0.91	**0.89**	0.72	0.86
discrete vision tokens							
Chameleon Team (2024)	-	-	-	-	-	-	0.39
Emu3-Gen Wang et al. (2024)	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Janus-Pro-7B Chen et al. (2025b)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Janus Wu et al. (2025)	0.97	0.68	0.30	0.84	0.46	0.42	0.61
show-o Xie et al. (2025)	0.98	0.80	0.66	0.84	0.31	0.50	0.68
show-o2 Xie et al. (2025)	0.99	0.86	0.55	0.86	0.46	0.63	0.73
MMaDA Yang et al. (2025b)	0.99	0.76	0.61	0.84	0.20	0.37	0.63
Lumina-DiMOO Xin et al. (2025)	**1.0**	**0.94**	**0.85**	0.89	0.85	**0.76**	**0.88**

Table 7: Performance comparison on the unified understanding and generation benchmark WISE. **Bold** indicates the best result and underline indicates the second best.

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
continuous vision tokens							
VILA-U-7b-256 Lin et al. (2023)	0.26	0.33	0.37	0.35	0.39	0.23	0.31
MetaQuery-XL Pan et al. (2025)	0.56	0.55	0.62	0.49	0.63	0.41	0.55
GPT-4o Hurst et al. (2024)	**0.81**	**0.71**	**0.89**	**0.83**	**0.79**	**0.74**	**0.80**
BAGEL Deng et al. (2025)	0.44	0.55	0.68	0.44	0.60	0.39	0.52
BAGEL w/ Self-CoT Deng et al. (2025)	0.76	0.69	0.75	0.65	0.75	0.58	0.70
discrete vision tokens							
Janus-1.3B Wu et al. (2025)	0.16	0.26	0.35	0.28	0.30	0.14	0.23
JanusFlow-1.3B Wu et al. (2025)	0.13	0.26	0.28	0.20	0.19	0.11	0.18
Janus-Pro-1B Wu et al. (2025)	0.20	0.28	0.45	0.24	0.32	0.16	0.26
Janus-Pro-7B Wu et al. (2025)	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Liquid Wu et al. (2024a)	0.38	0.42	0.53	0.36	0.47	0.30	0.41
Emu3 Wang et al. (2024)	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Orthus-7B-base Kou et al. (2024)	0.07	0.10	0.12	0.15	0.15	0.10	0.10
Orthus-7B-instruct Kou et al. (2024)	0.23	0.31	0.38	0.28	0.31	0.20	0.27
Show-o-demo Xie et al. (2025)	0.28	0.36	0.40	0.23	0.33	0.22	0.30

cluding Emu3-Gen, Janus-Pro-7B, show-o2, and Lumina-DiMOO, exhibit superior compositional generation and multi-object reasoning ability, with Lumina-DiMOO obtaining the highest overall score on GenEval.

Table 7 further compares unified models on the WISE Niu et al. (2025) benchmark. It is a unified benchmark that evaluates reasoning and generation within a single evaluation protocol. Models based on continuous vision tokens consistently outperform their discrete-token counterparts across most evaluation dimensions, particularly in spatial, temporal, and physics-related reasoning. Among them, GPT-4o achieves the best overall performance, indicating the strong advantage of tightly coupled reasoning and generation under continuous visual representations. In contrast, models relying on discrete vision tokens, such as the Janus and Orthus series, generally lag behind, suggesting that discretization may introduce representational bottlenecks that hinder joint reasoning–generation performance on WISE.

5.4 Challenges and Future Directions in Benchmarking

Despite the growing diversity of benchmarks, the evaluation of unified multimodal models remains fragmented. Existing benchmarks are often modality-imbalanced (neglecting audio or video), and automated metrics (e.g., FID, CLIP) fail to capture compositional correctness or factual grounding. Most benchmarks

still adopt static, single-turn tasks, whereas real-world unified agents require evaluation in dynamic, interactive, and long-horizon multimodal contexts. Future benchmarking should move toward open, adaptive frameworks that integrate human–AI co-evaluation, contextual reasoning, and unified scoring across perception, reasoning, and generation. Such comprehensive benchmarks will be key to tracking genuine progress toward general-purpose multimodal intelligence.

6 Challenges and opportunities

The development of unified multimodal models that integrate both understanding and generation remains in its infancy. While recent advances have demonstrated the feasibility of bridging autoregressive (AR) and diffusion paradigms, significant challenges persist across architectural, algorithmic, and data dimensions. In particular, it remains unclear whether multimodal understanding and generation constitute competing objectives that impose unavoidable trade-offs, or whether they can be jointly optimized in a synergistic manner as model capacity and data scale increase. At the same time, these challenges open new opportunities for innovation and cross-domain applications.

6.1 Whether Understanding and Generation Tasks Improve Each Other?

A central question in unified multimodal modeling is whether visual understanding and visual generation inherently compete for model capacity, or whether they can instead reinforce each other under appropriate design choices. Early multimodal systems often treated understanding and generation as loosely coupled or sequential processes, assuming that joint training would lead to negative interference. However, recent fully unified models Tong et al. (2025) Wu et al. (2024a) An et al. (2025) Zhu et al. (2023b) provide increasing evidence that understanding and generation are not independent capabilities, but can mutually reinforce each other when jointly optimized.

Liquid Wu et al. (2024a)reveal that a trade-off between language generation and visual generation does exist at smaller model scales, where limited representational capacity constrains simultaneous optimization. Notably, this trade-off is not fixed. As model size increases, the performance gap between unimodal and unified training progressively narrows, and in some cases nearly vanishes.

Liquid Wu et al. (2024a) further conduct controlled experiments, showing that adding understanding data consistently improves generation performance, while incorporating generation data also enhances understanding metrics under the same baseline. It attributes this phenomenon to the alignment of optimization objectives within a unified multimodal space. UniCTokens An et al. (2025) confirms this trend by systematically comparing unified models against understanding-only and generation-only counterparts, showing that unified training yields superior performance, particularly on tasks requiring semantic understanding and generation consistency.

Similarly, MetaMorph Tong et al. (2025) reports strong bidirectional gains: fixing the amount of generation data and increasing VQA data improves both understanding and generation, and vice versa. Notably, joint training significantly improves sample efficiency for generation, enabling models to acquire stable visual token generation with only a few thousand generation samples.

These results suggest that understanding and generation share substantial underlying structures, including vision–language alignment, semantic consistency constraints, and visual token formation mechanisms. Understanding tasks provide structured semantic supervision that stabilizes and constrains generation, while generation tasks force models to explicitly model fine-grained visual attributes and compositional details that are often underrepresented in understanding-only training. When optimized within a shared representation space, these two processes form a positive feedback loop, alleviating information sparsity and representation bias commonly observed in single-task training.

Significantly, MetaMorph further analyze the issue of whether certain visual understanding tasks correlate more strongly with generation performance. They reveals that tasks emphasizing general visual comprehension and vision-centric grounding exhibit the strongest correlation with visual generation quality, whereas knowledge-heavy understanding tasks contribute comparatively weaker gains. This finding indicates that

visual generation benefits primarily from understanding signals that reinforce perceptual alignment and grounded visual representations, rather than relying on abstract or symbolic reasoning alone.

Generally , these findings challenge the conventional view that understanding and generation are competing objectives. Instead, they support a unifying perspective in which understanding and generation become complementary processes under sufficient model capacity and aligned modality representations. From this standpoint, their mutual reinforcement is not an incidental artifact of training, but a natural consequence of unified multimodal modeling.

6.2 Challenges

Building upon the advances reviewed in the previous sections, this part focuses on the major challenges that remain unresolved in developing unified multimodal understanding and generation models. These challenges reveal the current limitations of existing methods and highlight areas where further innovation is required.

1) Visual Tokenization and Representation. A core difficulty lies in effectively representing visual information in a tokenized form suitable for autoregressive generation. Discrete tokenization methods based on VQ-VAE or VQ-GAN enable alignment with textual tokens but often sacrifice fine-grained visual details. Conversely, continuous representations preserve semantic richness but complicate sequence modeling and decoding. Achieving a unified tokenization scheme that balances efficiency, fidelity, and semantic expressiveness remains an open problem.

2) Architectural Divergence Between Understanding and Generation. Multimodal understanding models are primarily autoregressive, focusing on sequential reasoning and text prediction, whereas visual generation models rely heavily on diffusion processes for iterative denoising and synthesis. These fundamentally different modeling paradigms challenge architectural unification. Designing hybrid models that can seamlessly switch or jointly optimize both reasoning and generation remains a key research frontier.

3) Cross-Modal Alignment and Attention Mechanisms. Unified models must integrate heterogeneous modalities—text, image, video, and possibly audio—within a shared latent space. Scaling cross-modal attention to high-resolution visual inputs introduces substantial computational and memory overhead. Furthermore, maintaining semantic coherence across modalities, especially in complex tasks like instruction-based image generation or multimodal reasoning, poses an additional challenge.

4) Data Scarcity and Benchmark Limitations. Constructing large-scale, high-quality datasets that jointly support understanding and generation tasks is still difficult. Existing datasets are either biased toward single-modal tasks (e.g., captioning or VQA) or limited to generative tasks (e.g., text-to-image). Moreover, there is a lack of standardized benchmarks that simultaneously evaluate reasoning accuracy, visual quality, and cross-modal consistency. Comprehensive evaluation frameworks are urgently needed.

5) Computational Cost and Scalability. Unified multimodal models require vast computational resources due to their multimodal encoders, diffusion decoders, and autoregressive backbones. Training such large-scale systems demands extensive data and compute budgets, which limits accessibility and reproducibility. Efficient architectures, model compression, and modular training pipelines are essential for scalable deployment.

6.3 Opportunities

Despite these challenges, the pursuit of unification offers remarkable opportunities for advancing multimodal AI research and applications.

1) Toward General-Purpose Multimodal Intelligence. A successful unified model can both *understand* and *generate* multimodal content, enabling end-to-end perception and creation within a single framework. Such models could serve as general-purpose agents capable of reasoning over complex visual scenes and producing coherent, high-quality visual outputs in response.

2) Emergent Capabilities Through Modality Fusion. Unifying understanding and generation may yield emergent capabilities that neither paradigm can achieve alone. For instance, reasoning-augmented

image generation can improve semantic controllability, while generation-enhanced understanding can enable visual imagination and explanation in natural language reasoning tasks.

3) Architectural Innovation. The coexistence of autoregressive and diffusion mechanisms encourages the exploration of new hybrid frameworks that leverage the strengths of both paradigms. Promising directions include bidirectional AR-diffusion coupling, shared latent token spaces, and unified decoders capable of multimodal synthesis.

4) Advancement of Datasets and Evaluation Protocols. The unification trend motivates the construction of novel datasets encompassing both understanding and generation signals, as well as comprehensive benchmarks evaluating semantic accuracy, visual quality, and reasoning consistency. This will foster more standardized and comparable research progress.

5) Practical Applications and Ecosystem Growth. Unified models are poised to power next-generation applications in interactive AI, visual storytelling, robotics, design, and education. Their ability to reason and generate across modalities will enable seamless multimodal interaction and content creation, expanding the reach of AI systems into creative and scientific domains.

6.4 Summary

In summary, while unified multimodal understanding and generation models face nontrivial challenges in representation, architecture, data, and scalability, they also present transformative opportunities for the future of AI. Continued exploration of hybrid architectures, efficient training strategies, and standardized evaluation will accelerate progress toward truly general multimodal intelligence.

References

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

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. *International Conference on Computer Vision*, pp. 8947–8956, 2019. URL `https://api.semanticscholar.org/CorpusID:56517630`.

Liefu Ai, Junqing Yu, Zebin Wu, Yunfeng He, and Tao Guan. Optimized residual vector quantization for efficient approximate nearest neighbor search. *Multimedia Systems*, 23(2):169–181, 2017.

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

Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. Openleaf: Open-domain interleaved image-text generation and evaluation. *arXiv preprint arXiv:2310.07749*, 2023.

Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, Bocheng Zou, Chaoqun Yang, and Wentao Zhang. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *ArXiv*, abs/2505.14671, 2025. URL `https://api.semanticscholar.org/CorpusID:278769842`.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Artem Babenko and Victor Lempitsky. Additive quantization for extreme vector compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 931–938, 2014.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Anton Bakhtin, Arthur Szlam, and Marc'Aurelio Ranzato. Geneval: A benchmark suite for evaluating generative models. 2018. URL `https://api.semanticscholar.org/CorpusID:53351804`.

Christopher F Barnes, Syed A Rizvi, and Nasser M Nasrabadi. Advances in residual vector quantization: A review. *IEEE transactions on image processing*, 5(2):226–262, 1996.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

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

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

M Caron, H Touvron, I Misra, H Jégou, J Mairal, P Bojanowski, and A Joulin. Self-supervised vision transformers with dino. *GitHub Repository. Available online: https://github. com/facebookresearch/dino (accessed on 11 August 2022)*.

Di Chang, Mingdeng Cao, Yichun Shi, Bo Liu, Shengqu Cai, Shijie Zhou, Weilin Huang, Gordon Wetzstein, Mohammad Soleymani, and Peng Wang. Bytemorph: Benchmarking instruction-guided image editing with non-rigid motions. *arXiv preprint arXiv:2506.03107*, 2025.

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

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

Yongjian Chen, Tao Guan, and Cheng Wang. Approximate nearest neighbor search by residual vector quantization. *Sensors*, 10(12):11259–11273, 2010.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Zisheng Chen, Chunwei Wang, Xiuwei Chen, Hongbin Xu, Runhui Huang, Jun Zhou, Jianhua Han, Hang Xu, and Xiaodan Liang. Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. *arXiv preprint arXiv:2503.06764*, 2025c.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024.

Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, et al. On path to multimodal generalist: General-level and general-bench. In *Forty-second International Conference on Machine Learning*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. URL `https://api.semanticscholar.org/CorpusID:259243928`.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Tanmay Gautam, Reid Pryzant, Ziyi Yang, Chenguang Zhu, and Somayeh Sojoudi. Soft convex quantization: Revisiting vector quantization with convex optimization. *arXiv preprint arXiv:2310.03004*, 2023.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2946–2953, 2013.

Y Ge, Y Ge, Z Zeng, X Wang, and Y Shan. Planting a seed of vision in large language model. arxiv 2023. *arXiv preprint arXiv:2307.08041*.

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

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Iris AM Huijben, Matthijs Douze, Matthew Muckley, Ruud JG Van Sloun, and Jakob Verbeek. Residual quantization with implicit neural codebooks. *arXiv preprint arXiv:2401.14732*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

Yannis Kalantidis and Yannis Avrithis. Locally optimized product quantization for approximate nearest neighbor search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2321–2328, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Benjamin Klein and Lior Wolf. End-to-end supervised product quantization for image search and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5041–5050, 2019.

Siqi Kou, Jiachun Jin, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *ArXiv*, abs/2412.00127, 2024. URL `https://api.semanticscholar.org/CorpusID:274437723`.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36: 71683–71702, 2023.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024b.

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

Jindong Li, Yali Fu, Jiahong Liu, Linxiao Cao, Wei Ji, Menglin Yang, Irwin King, and Ming-Hsuan Yang. Discrete tokenization for multimodal llms: A comprehensive survey. *arXiv preprint arXiv:2507.22920*, 2025a.

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

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024c.

Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6955–6965, 2024d.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*, 2025b.

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

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

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*, 2024f.

Huawei Lin, Tong Geng, Zhaozhuo Xu, and Weijie Zhao. Vtbench: Evaluating visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2505.13439*, 2025.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26679–26689, 2023. URL `https://api.semanticscholar.org/CorpusID:266174746`.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*, 2024.

Shicong Liu, Hongtao Lu, and Junru Shao. Improved residual vector quantization for high-dimensional approximate nearest neighbor search. *arXiv preprint arXiv:1509.05195*, 2015.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2023. URL `https://api.semanticscholar.org/CorpusID:259837088`.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023a.

Zepu Lu, Defu Lian, Jin Zhang, Zaixi Zhang, Chao Feng, Hao Wang, and Enhong Chen. Differentiable optimized product quantization and beyond. In *Proceedings of the ACM Web Conference 2023*, pp. 3353–3363, 2023b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Julieta Martinez, Holger H Hoos, and James J Little. Stacked quantizers for compositional vector compression. *arXiv preprint arXiv:1411.2173*, 2014.

Julieta Martinez, Joris Clement, Holger H Hoos, and James J Little. Revisiting additive quantization. In *European Conference on Computer Vision*, pp. 137–153. Springer, 2016.

Yusuke Matsui, Yusuke Uchida, Hervé Jégou, and Shin'ichi Satoh. A survey of product quantization. *ITE Transactions on Media Technology and Applications*, 6(1):2–10, 2018.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kun-Peng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *ArXiv*, abs/2503.07265, 2025. URL `https://api.semanticscholar.org/CorpusID:276929205`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023.

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

Yatian Pang, Peng Jin, Shuo Yang, Bin Lin, Bin Zhu, Zhenyu Tang, Liuhan Chen, Francis EH Tay, Ser-Nam Lim, Harry Yang, et al. Next patch prediction for autoregressive visual generation. *arXiv preprint arXiv:2412.15321*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Zhang Shengyu, Dong Linfeng, Li Xiaoya, Zhang Sen, Sun Xiaofei, Wang Shuhe, Li Jiwei, Runyi Hu, Zhang Tianwei, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL `https://arxiv.org/abs/2508.10104`.

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

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023a.

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

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Yuhta Takida, Takashi Shibuya, WeiHsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. *arXiv preprint arXiv:2205.07547*, 2022.

Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv preprint arXiv:2503.21193*, 2025.

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

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

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

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17001–17012, 2025.

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

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

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

Ivan Volkov. Homology-constrained vector quantization entropy regularizer. *arXiv preprint arXiv:2211.14363*, 2022.

Kai Wang, Dongwen Tang, Wangbo Zhao, Konstantin Schürholt, Zhangyang Wang, and Yang You. Recurrent diffusion for large-scale parameter generation. *arXiv preprint arXiv:2501.11587*, 2025a.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020.

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

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv e-prints*, pp. arXiv–2412, 2024a.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

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

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Donna Xu, Ivor W Tsang, and Ying Zhang. Online product quantization. *IEEE Transactions on Knowledge and Data Engineering*, 30(11):2185–2198, 2018.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025b.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. URL `https://api.semanticscholar.org/CorpusID:3104920`.

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

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

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2023. URL `https://api.semanticscholar.org/CorpusID:265466525`.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-

$$\backslash$$

alpha : $Large - scaleautoregressivepretrainedchineselanguagemodelswithauto - parallelcomputation.$ *arXiv preprint arXiv:2104.12369*, 2021.

Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. *arXiv preprint arXiv:2505.01490*, 2025a.

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

Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18467–18476, 2023a.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023b.

Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kai-Ni Wang, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, Bin Cui, et al. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 37:96963–96992, 2024.

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

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

Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22798–22807, 2023.

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

Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 56–66, 2025.

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

Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv preprint arXiv:2312.09251*, 2023b.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023c.

Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models. *arXiv preprint arXiv:2503.08686*, 2025.