

VoT: Vision-of-Thought for Unified Multimodal Representation Alignment

Jingxiang Sun*, Chao Liao*[†], Zhengxiong Luo*, Chaorui Deng, Chen-lin Zhang,
Junke Wang, Ceyuan Yang, Haoqi Fan, Weilin Huang[‡]

ByteDance Seed

*Equal Contribution, [†]Project Lead, [‡]Corresponding Author

Abstract

Current text-to-image systems typically employ a “text encoder plus diffusion decoder” paradigm, in which text semantics directly modulate continuous latent noise. Despite their success, these methods lack an explicit, interpretable intermediate representation that effectively bridges high-level linguistic semantics and low-level visual signals. In this paper, we propose Vision-of-Thought (VoT), a framework that introduces a discrete visual-thinking layer between vision-language models (VLMs) and diffusion transformers (DiTs). Instead of treating VLMs merely as text encoders, we use them as multimodal planners that generate discrete VoT tokens representing high-level visual plans, such as objects and layouts, before rendering pixels. We train a specialized VoT tokenizer in the VLM semantic space with a closed-loop objective that combines VLM alignment, feature reconstruction, and vector-quantization losses. These objectives make the tokens semantically readable by the VLM while preserving the visual information needed for generation. Experimental results demonstrate that VoT improves semantic alignment and provides a structured interface for interpretable and controllable generation.

Date: September 9, 2026

1 Introduction

A unified model capable of both deep understanding and high-fidelity generation holds immense potential for the future of AI. Such a system could seamlessly generate images based on complex instructions, reason about visual data, and visualize multimodal analyses through generated outputs. The unveiling of GPT-4o’s enhanced capabilities has further highlighted this potential, sparking widespread interest in unification.

However, designing such a unified framework presents a formidable challenge. It requires integrating the specific strengths of autoregressive models—which excel at discrete reasoning and text generation—with the robustness of diffusion-based models for high-quality image synthesis. Current text-to-image (T2I) paradigms typically treat this as a direct mapping problem, encoding text prompts into static embeddings or key-value (KV) caches that condition a diffusion process. While effective for simple descriptions, this approach faces a significant modality gap. The rich, structured world knowledge embedded in a large language model (LLM) is difficult to transfer effectively to the diffusion decoder through static embeddings alone, leading to a bottleneck where the model “knows” what to generate but fails to “render” it accurately.

In this work, we introduce Vision-of-Thought (VoT) as a semantic bridge to align these distinct modalities.



Figure 1 Images generated by VoT from complex text and image prompts.

Adopting a “parallel expert” philosophy, we freeze a powerful vision-language model (VLM) to preserve its multimodal understanding capabilities and introduce lightweight autoregressive experts to generate discrete semantic visual tokens. Unlike recent query-based approaches (e.g., MetaQuery [22], BLIP-3o [33]) that generate parallel queries for conditioning, VoT generates visual tokens autoregressively. This choice is deliberate: while query-based methods offer parallel efficiency, they lack the structural consistency with VLMs required for complex reasoning. By treating visual generation as a next-token prediction task, VoT allows the model to perform “visual chain-of-thought,” planning the image composition step by step before rendering—effectively extending the VLM’s reasoning capability into the visual domain.

A critical hurdle for autoregressive (AR) image generation is visual tokenization. Standard visual tokenizers (e.g., VQ-VAE, VQGAN) are optimized primarily for reconstruction fidelity, compressing images into “pixel codes” dominated by local textures and low-level details. As a result, their codebooks are fundamentally misaligned with the semantic representation space of VLMs: token sequences are often excessively long and semantically redundant, behaving more like compressed pixel fragments than meaningful semantic units. Moreover, these reconstruction-centric tokens tend to exhibit unfavorable statistics for next-token prediction (e.g., heavy-tailed usage and weak semantic correlation among neighbors), making AR modeling unstable. Consequently, when a VLM is tasked with predicting such tokens, it often struggles to maintain semantic coherence, which degrades downstream generation quality.

To address this mismatch, we propose a *VoT Tokenizer* that is explicitly aligned with the VLM semantic space through a *closed-loop alignment* paradigm. Instead of quantizing raw pixels, we perform *semantic quantization* by using the frozen ViT features of a pretrained VLM as the quantization target, ensuring discrete tokens originate from a high-level semantic manifold. We optimize the tokenizer with three objectives: (i) a *VLM alignment loss* that enables the VLM to “read” and use the quantized tokens, (ii) a *feature reconstruction loss* that recovers the teacher ViT feature space, and (iii) a *vector-quantization loss* that stabilizes codebook

learning. The first two objectives are strongly complementary: VLM alignment provides language-grounded supervision, while reconstruction supplies dense, information-preserving constraints. Furthermore, we observe that *scaling the teacher VLM* consistently improves tokenizer quality and downstream semantic alignment, suggesting that stronger teachers provide cleaner and more compositional supervision.

On the model side, we introduce an efficient architecture that integrates VoT into modern unified models by extending *Mixture-of-Transformers (MoT)*. Specifically, we initialize a VoT branch by copying a subset of pretrained VLM experts and keep the original VLM branch frozen throughout training. This design preserves the VLM’s multimodal understanding capability while enabling efficient visual-planning training with minimal additional compute. After tokenizer training, our model training proceeds in two stages: (i) *VoT branch pretraining*, where the model learns to produce structured VoT tokens from text–image pairs, and (ii) *VoT-DiT joint training*, where VoT prediction and diffusion-based rendering are jointly optimized so that the tokens serve as explicit visual plans.

Our contributions are summarized as follows:

- **Vision-of-Thought Framework.** We propose *VoT*, a framework that inserts a discrete visual-thought branch between a VLM and a diffusion decoder, decoupling semantic reasoning from pixel-space rendering and improving cross-modal semantic alignment.
- **VLM-Aligned VoT Tokenizer.** We introduce a tokenizer training recipe that quantizes *VLM semantic features* rather than pixels and uses a closed-loop objective combining VLM alignment, feature reconstruction, and vector-quantization losses. We show that VLM alignment and reconstruction are complementary and that *scaling the teacher VLM* consistently benefits token quality and downstream alignment.
- **Efficient Unified MoT Integration and Strong Results.** We present a practical MoT-based instantiation of VoT by cloning VLM experts to initialize a VoT branch and freezing the VLM branch for efficient training and stable understanding. With a two-stage model-training pipeline after Stage 0 tokenizer training, using only text–image pairs throughout, we achieve substantial improvements in semantic alignment and reasoning-driven text-to-image generation.

2 Related Work

2.1 Unified Multimodal Understanding and Generation

The convergence of visual understanding and generation into unified architectures represents a pivotal frontier in AI research. Recent breakthroughs, particularly GPT-4o [21], have demonstrated the potential of single models that can seamlessly interpret and generate multimodal content. Along this trajectory, various approaches have emerged to unify these capabilities. **Unified autoregressive models** such as Show-o [32], Janus [5], and Unified-IO [18] attempt to handle both tasks within a single transformer architecture by treating images as discrete token sequences. However, these methods often compromise either generation fidelity or understanding depth because of the tension between reconstruction-oriented visual codes and semantics-oriented language modeling.

Alternatively, **diffusion-based unification efforts** [11] typically graft understanding modules onto diffusion backbones, yet they suffer from a modality gap: the rich, structured knowledge of VLMs cannot be transferred effectively to diffusion decoders through static conditioning alone. VoT introduces an explicit, interpretable intermediate layer—discrete visual thoughts—that decouples semantic planning from pixel rendering, preserving the specialized strengths of both VLMs and diffusion transformers.

2.2 Visual Tokenization Strategies

Visual tokenization serves as the critical interface between continuous visual signals and discrete language-like representations. Early works such as VQ-VAE [19] and VQGAN [8] established the paradigm of compressing images into discrete tokens via reconstruction objectives. Subsequent methods such as LFQ [34] and MaskGIT [2] improved codebook efficiency and generation diversity. However, these tokenizers remain

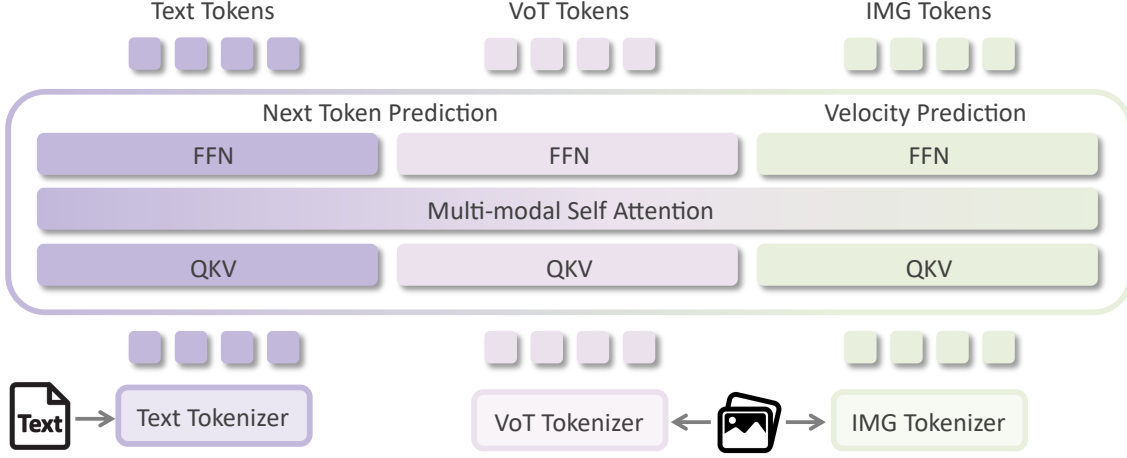


Figure 2 Architecture of the unified vision-language model with VoT. We extend MoT to three branches: *Text*, *VoT*, and *Diffusion*. Key-value pairs from all branches are concatenated for unified attention. For text-to-image and image-to-image generation, the VoT branch first predicts discrete visual tokens as semantic plans; the diffusion branch then attends jointly to the text and VoT key-value pairs to synthesize the final image.

fundamentally *reconstruction-centric*: they optimize for pixel-level fidelity, producing “compressed pixel fragments” rather than semantic units.

Recent efforts to align visual tokens with language semantics have explored alternative pathways. SEED [11] and LaVIT [14] align visual tokens with language-model embeddings but focus primarily on post-hoc alignment rather than native integration into the VLM’s reasoning process. Emu [28] introduces a visual decoder for VLM-generated embeddings but relies on continuous rather than discrete representations. Across these approaches, the key limitation is the lack of closed-loop training that preserves visual information while making discrete tokens semantically readable by a VLM. Our VoT tokenizer addresses this gap with a unified objective that combines VLM alignment, feature reconstruction, and vector-quantization losses.

2.3 Reasoning and Planning in Visual Generation

The integration of reasoning capabilities into generative models has gained significant traction. Text-to-image systems such as DALL-E 3 [1] and Imagen [27] improve prompt following through sophisticated text encoding yet lack explicit intermediate reasoning steps. Recent efforts to introduce chain-of-thought (CoT) mechanisms into vision tasks [6, 35] have primarily focused on understanding rather than generation.

Several concurrent works explore planning-based generation. LayoutGPT [10] leverages LLMs for layout planning before image synthesis. However, these methods typically treat planning and generation as separate stages. VoT integrates planning *within* the generative process through a unified token vocabulary. By generating discrete visual plans as VoT tokens, our framework enables the VLM to perform “visual thinking”—reasoning about composition, spatial relationships, and attributes before committing to pixel values. This shifts the paradigm from direct *text-to-image mapping* to *text-to-thought-to-image generation*, providing interpretability and controllability that are absent from end-to-end diffusion or query-based approaches.

3 Overview of VoT

Our proposed *Vision-of-Thought (VoT)* framework bridges the gap between semantic understanding and visual synthesis by inserting a discrete visual-thought layer between a VLM and a diffusion decoder. Rather than mapping text directly to pixels, the VLM first produces structured VoT tokens that act as explicit visual plans, which are then rendered by the diffusion branch in the unified model (Fig. 2).

The framework follows three stages. In *Stage 0: VLM-Aligned VoT Tokenizer Training* (Sec. 4, Fig. 3), we

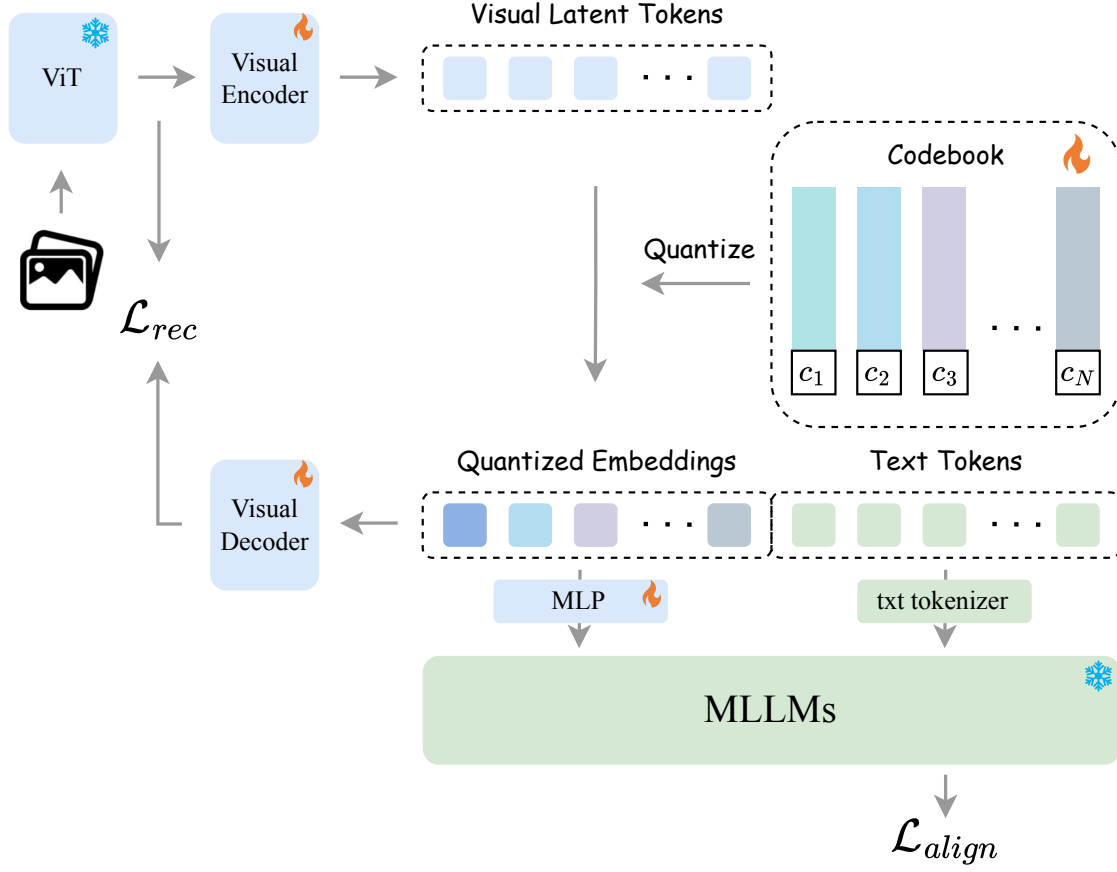


Figure 3 VLM-aligned VoT tokenizer training pipeline. Images are encoded by a frozen VLM ViT, adapted by a trainable encoder, and discretized with SimVQ. Training combines (1) feature reconstruction loss $\mathcal{L}_{rec}$, (2) VLM alignment loss $\mathcal{L}_{align}$, and (3) vector-quantization loss $\mathcal{L}_{VQ}$.

learn a discrete codebook in the VLM semantic space using feature reconstruction, VLM alignment, and vector-quantization losses. In *Stage 1: VoT Branch Pretraining* (Sec. 5), we freeze the pretrained VLM and diffusion branches and train lightweight VoT experts to predict visual tokens from text prompts using targets extracted from paired images. In *Stage 2: VoT-DiT Joint Training* (Sec. 5), we jointly fine-tune the VoT and diffusion branches while keeping the VLM branch frozen to preserve multimodal understanding and reduce training cost.

4 VLM-Aligned VoT Tokenizer Training

The cornerstone of our framework is a discrete visual representation natively compatible with a vision-language model (VLM). Unlike traditional tokenizers prioritizing pixel-level fidelity, our *VoT Tokenizer* minimizes the distributional shift between visual tokens and the VLM’s semantic space via a specialized architecture and a unified alignment objective.

4.1 Semantically Aligned Tokenizer Architecture

Our architecture (Fig. 3) is designed to fully harness the pretrained capabilities of state-of-the-art VLMs. The pipeline comprises three primary modules:

Foundation & Adapter Encoder. The encoder integrates a frozen foundation model with a trainable adapter. We denote the Vision Transformer (ViT) of the target VLM (Qwen2.5-VL) as $\mathcal{E}_{\text{frozen}}$, keeping it **frozen** to preserve its robust, pre-aligned semantic features. To bridge the modality gap between continuous ViT features and the discrete latent space, we append a lightweight trainable adapter $\mathcal{E}_{\text{adapt}}$ (a 1-layer ViT block, denoted as *Visual Encoder* in Fig. 3) to the backbone. Given an input image $\mathbf{x}$, adapted semantic features $\mathbf{h}$ are derived via:

$$\mathbf{h} = \mathcal{E}_{\text{adapt}}(\mathcal{E}_{\text{frozen}}(\mathbf{x})). \quad (1)$$

This adapter projects foundation features onto a manifold suitable for quantization while retaining the VLM’s semantic structure.

SimVQ Head. We employ a Similarity-based Vector Quantization (SimVQ) mechanism with a learnable codebook $\mathcal{C} = \{c_k\}_{k=1}^K \subset \mathbb{R}^d$. The discrete code index k for a feature vector $\mathbf{h}_i$ is selected by minimizing the Euclidean distance:

$$k = \arg \min_{j \in \{1, \dots, K\}} \|\mathbf{h}_i - c_j\|_2^2. \quad (2)$$

To prevent codebook collapse, we regularize the similarity logits $s_{i,j} = -\|\mathbf{h}_i - c_j\|_2^2$ with an **entropy loss**. The quantized feature vector $\mathbf{z}_q$ is passed to subsequent layers, with gradients propagated via the straight-through estimator: $\mathbf{z}_{st} = \mathbf{h} + \text{sg}[\mathbf{z}_q - \mathbf{h}]$, where $\text{sg}[\cdot]$ denotes the stop-gradient operator.

Deep Semantic Decoder. The decoder $\mathcal{D}$ reconstructs continuous semantic features from quantized tokens $\mathbf{z}_{st}$ via $\hat{\mathbf{f}} = \mathcal{D}(\mathbf{z}_{st})$. Departing from typical lightweight designs, we employ a 6-layer transformer decoder. As shown in Table 2, increasing the decoder depth from 2 layers in Exp.D (69.32) to 6 layers in Exp.E (70.37) consistently improves AutoEval. We attribute this gain to the decoder’s increased capacity to *invert quantization* in the teacher feature space: a stronger decoder reduces the pressure on the codebook to preserve fine-grained details, allowing tokens to focus on language-grounded semantics while the decoder smooths quantization noise when reconstructing dense ViT features.

4.2 Training Objectives

We train the VoT tokenizer with three complementary objectives as follows. Throughout training, the teacher ViT encoder and the VLM are kept frozen.

Feature Reconstruction Loss ($\mathcal{L}_{rec}$): Visual Grounding. Given an image $\mathbf{x}$, the frozen teacher ViT produces token features $\mathbf{f} = \mathcal{E}_{\text{frozen}}(\mathbf{x}) \in \mathbb{R}^{N \times d}$. After quantization and straight-through estimation, we obtain $\mathbf{z}_{st}$, and a semantic decoder $\mathcal{D}$ reconstructs teacher features $\hat{\mathbf{f}} = \mathcal{D}(\mathbf{z}_{st})$. We use a token-wise mean-squared error:

$$\mathcal{L}_{rec} = \frac{1}{N} \left\| \mathbf{f} - \hat{\mathbf{f}} \right\|_2^2. \quad (3)$$

VLM Alignment Loss ($\mathcal{L}_{align}$): Semantic Readability. This objective enforces that the discretized representation remains readable by the frozen VLM. We feed the tokenizer-produced visual embeddings (denoted as $\mathbf{v}$; e.g., $\mathbf{v} = \mathbf{z}_{st}$ or a projected version thereof) into the frozen VLM and minimize the standard language-model cross-entropy on the paired caption $\mathbf{y} = (y_1, \dots, y_T)$:

$$\mathcal{L}_{align} = - \sum_{t=1}^T \log P_{\text{VLM}}(y_t \mid y_{<t}, \mathbf{v}). \quad (4)$$

Vector-Quantization Loss ($\mathcal{L}_{VQ}$): Stable Quantization. Let $\mathbf{h}$ denote the pre-quantization features (the tokenizer output before nearest-neighbor lookup), and $\mathbf{z}_q$ denote the selected codebook embeddings. We use the standard VQ objective (codebook loss + commitment loss):

$$\mathcal{L}_{VQ} = \|\text{sg}[\mathbf{h}] - \mathbf{z}_q\|_2^2 + \beta \|\mathbf{h} - \text{sg}[\mathbf{z}_q]\|_2^2, \quad (5)$$

where $\text{sg}[\cdot]$ is the stop-gradient operator and β is the commitment weight.

Total Objective. The full tokenizer training objective is:

$$\mathcal{L}_{tok} = \mathcal{L}_{rec} + \lambda_{align}\mathcal{L}_{align} + \lambda_{VQ}\mathcal{L}_{VQ}. \quad (6)$$

5 Unified Vision-Language Model with VoT

Building upon our pretrained visual tokenizer, we now introduce a unified multimodal architecture that integrates discrete visual tokens into a joint understanding and generation framework. Our approach extends the Mixture-of-Transformers (MoT) paradigm to support three specialized pathways: multimodal understanding, visual token prediction, and diffusion-based image synthesis.

5.1 Unified Three-Branch MoT Backbone

Building upon the Mixture-of-Transformers (MoT) paradigm [7, 16], which harmonizes language and diffusion generation within a shared backbone, we extend the architecture to a three-branch design by incorporating a dedicated *VoT branch* for discrete visual token modeling. In this unified framework, distinct modalities share the transformer backbone while using modality-specific components to preserve their unique properties. Specifically, for each transformer layer ℓ , we maintain separate hidden states for text ($\mathbf{h}_{\text{txt}}^{(\ell)}$), visual tokens ($\mathbf{h}_{\text{vot}}^{(\ell)}$), and diffusion latents ($\mathbf{h}_{\text{diff}}^{(\ell)}$). Each branch $m \in \{\text{txt}, \text{vot}, \text{diff}\}$ applies independent layer normalization (LN) and projections to compute queries, keys, and values:

$$\begin{aligned} \mathbf{Q}_m &= \mathbf{W}_m^Q \cdot \text{LN}_m(\mathbf{h}_m^{(\ell)}), \\ \mathbf{K}_m &= \mathbf{W}_m^K \cdot \text{LN}_m(\mathbf{h}_m^{(\ell)}), \\ \mathbf{V}_m &= \mathbf{W}_m^V \cdot \text{LN}_m(\mathbf{h}_m^{(\ell)}). \end{aligned} \quad (7)$$

To facilitate cross-modal information exchange, we construct a global context by concatenating the key-value pairs from all branches:

$$\mathbf{K} = [\mathbf{K}_{\text{txt}}; \mathbf{K}_{\text{vot}}; \mathbf{K}_{\text{diff}}], \quad \mathbf{V} = [\mathbf{V}_{\text{txt}}; \mathbf{V}_{\text{vot}}; \mathbf{V}_{\text{diff}}]. \quad (8)$$

The unified attention mechanism then operates over this joint space, modulated by a branch-specific mask $\mathbf{M}_m$:

$$\text{Attn}_m = \text{softmax}\left(\frac{\mathbf{Q}_m \mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}_m\right) \mathbf{V}. \quad (9)$$

Crucially, we employ causal masking for the text and VoT branches to enable autoregressive generation, whereas the diffusion branch uses bidirectional attention. Finally, the layer output is obtained by processing the attention result through a branch-specific feed-forward network (FFN):

$$\mathbf{h}_m^{(\ell+1)} = \mathbf{h}_m^{(\ell)} + \mathbf{W}_m^O \cdot \text{Attn}_m + \text{FFN}_m(\text{LN}'_m(\cdot)). \quad (10)$$

This architecture balances shared semantic modeling with modality-specific processing.

We implement VoT by adding per-layer VoT experts and a codebook-sized embedding and prediction interface for visual tokens, while the diffusion branch predicts denoising velocities. To integrate VoT efficiently into a unified MoT backbone, we initialize the VoT experts by copying a subset of pretrained VLM experts and keep the VLM branch frozen to preserve multimodal understanding and reduce training cost. At inference time, we first generate VoT tokens autoregressively conditioned on text, then render the final image with the diffusion branch conditioned on both text and VoT KV caches.

5.2 Training Recipe

Following Stage 0 tokenizer training, we adopt a two-stage strategy to integrate the VoT branch into the pretrained Mogao model.

Stage 1: VoT Branch Pretraining. In this stage, we warm up the newly introduced VoT branch while freezing the pretrained backbone, so that VoT learns to represent and predict discrete visual tokens without perturbing the model’s existing understanding and rendering capabilities.

We initialize the backbone from *Mogao* [16], a 14B MoT-based unified model. We load the pretrained weights for the VLM and diffusion branches and initialize the per-layer VoT experts by copying a subset of pretrained VLM experts. Only the remaining VoT-specific components are randomly initialized (std. 0.0125): the visual token embedding table $\mathbf{E}_{\text{vot}} \in \mathbb{R}^{65,536 \times 2,048}$ and the VoT prediction head. We use the frozen visual tokenizer from Stage 0 (a Qwen2.5-VL ViT encoder followed by an adapter encoder and a SimVQ quantizer; Sec. 4) to extract discrete supervision targets and keep it fixed throughout this stage. The training objective is an autoregressive cross-entropy loss over a 65,536-way codebook, predicting 1,024 VoT tokens per image.

Stage 2: VoT-DiT Joint Training. Starting from the Stage 1 checkpoint, we jointly train the *VoT branch* and the *diffusion branch* so that they can co-adapt for high-quality synthesis. Concretely, we initialize from the warmed-up VoT model and update parameters in both pathways, including VoT experts, diffusion experts, and shared components. This joint optimization allows the diffusion branch to better exploit VoT conditioning, while the VoT branch refines its token representations to be more useful for generation.

We optimize a combined objective consisting of the *VoT autoregressive prediction loss* and the *diffusion velocity prediction loss*. Unlike Stage 1, where gradients are restricted to VoT modules, the diffusion loss now backpropagates through the shared attention layers and updates both branches, tightening the coupling between token prediction and rendering quality.

To support classifier-free guidance at inference time, we apply structured conditional dropout during training: VoT tokens are dropped with probability 0.5, and text tokens are dropped with probability 0.1 when VoT is present, resulting in approximately 50% text-only, 45% text-plus-VoT, and 5% VoT-only samples. In addition, we randomly mask 0–20% of VoT key–value pairs during attention, which prevents the diffusion branch from over-relying on VoT and improves robustness to incomplete conditions.

6 Experiments

6.1 Implementation Details

Data Configuration. In Stage 0, we use image–text pairs with images processed at native aspect ratios and resolutions from approximately 434×434 to 1024×1024 pixels (188K–1M total pixels). We set the maximum sequence length to 16,384 tokens, with a per-sample limit of 12,288 tokens. Stages 1 and 2 use a large-scale text-to-image dataset. Additional dataset statistics and training configurations are provided in the Appendix.

Model Configuration. We use the Qwen2.5-VL-7B ViT as the frozen vision encoder for the visual tokenizer. The SimVQ quantizer comprises a 1-layer encoder, a codebook with 65,536 entries ($d = 128$), and a 6-layer decoder. We train it with a learning rate of 3×10^{-4} using vector-quantization, feature reconstruction, and VLM alignment losses. For the unified vision-language model, we initialize from the pretrained Mogao-14B checkpoint and add a dedicated 7B-parameter VoT branch, yielding 21B parameters in total.

6.2 Text-to-Image Generation

Evaluation Setup. We evaluate text-to-image generation on the public GenEval benchmark [13] and an internal benchmark. GenEval measures compositional text-to-image alignment with object detectors, covering object categories, counts, colors, and spatial relations. It contains 553 prompts, and we generate four images per prompt. We also evaluate on our internal DreamBench suite with an automatic pipeline and report its aggregate metric as AutoEval.

Quantitative Comparison. We report GenEval results in Table 1. VoT achieves an overall score of 0.91, outperforming leading generation-specialized models (e.g., FLUX.1-dev, 0.82) and recent unified multimodal models (e.g., BAGEL, 0.82; Mogao, 0.89). VoT also shows clear advantages on compositionally demanding

Table 1 Quantitative evaluation on GenEval benchmark. We compare VoT with state-of-the-art “Gen. Only” (generation-specialized) and “Unified” (understanding + generation) models. † denotes methods utilizing an LLM rewriter. **VoT** achieves the highest overall score on GenEval, showing significant advantages in spatial and counting tasks.

Type	Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
<i>Gen. Only</i>	PixArt- α [3]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [26]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [25]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [31]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [23]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [20]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [9]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
	FLUX.1-dev† [15]	0.98	0.93	0.75	0.93	0.68	0.65	0.82
<i>Unified</i>	Chameleon [29]	–	–	–	–	–	–	0.39
	LWM [17]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X [12]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	TokenFlow-XL [24]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	ILLUME [30]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
	Janus [5]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Transfusion [36]	–	–	–	–	–	–	0.63
	Emu3-Gen† [31]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
	Show-o [32]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
	Janus-Pro-7B [4]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
	MetaQuery-XL† [22]	–	–	–	–	–	–	0.80
	BAGEL [7]	0.99	0.94	0.81	0.88	0.64	0.63	0.82
	Mogao [16]	1.00	0.97	0.83	0.93	0.84	0.80	0.89
	VoT (Ours)	1.00	0.98	0.86	0.95	0.85	0.82	0.91

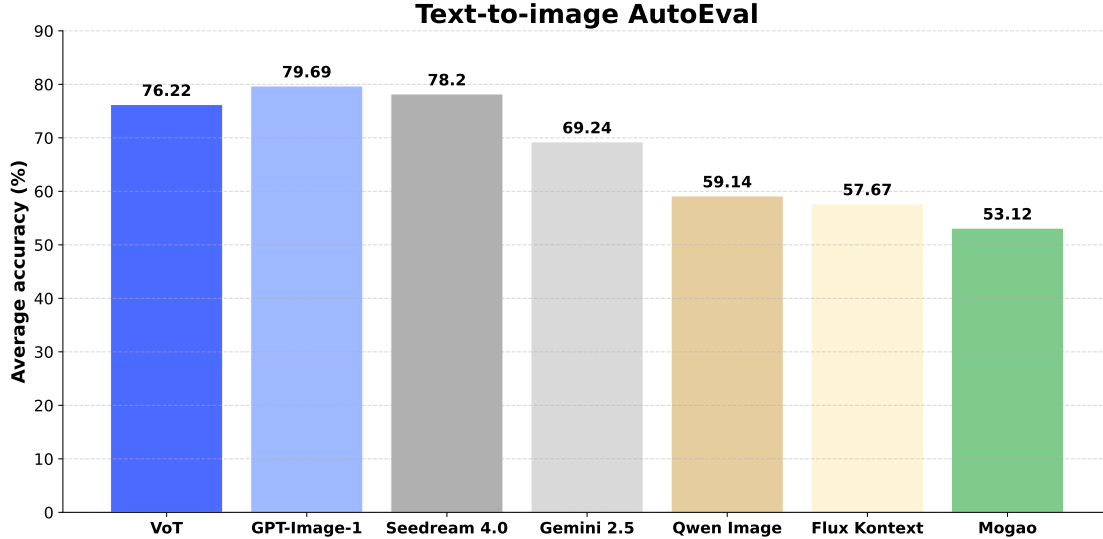


Figure 4 Evaluation of text-to-image instruction alignment.

tasks such as counting (0.86 vs. 0.83 for Mogao). These results support our hypothesis that the explicit discrete VoT layer transfers the VLM’s planning and reasoning capabilities to the generation process, helping satisfy structural constraints that implicit conditioning methods often fail to handle.

On our internal instruction-following benchmark (Fig. 4), VoT achieves strong prompt–image coherence and outperforms baselines such as Seedream 4.0 and GPT-Image-1, demonstrating the benefit of VoT tokens for instruction adherence.



Figure 5 Text-to-image comparison against GPT-Image-1. For each pair of images, left: VoT; right: GPT-Image-1.

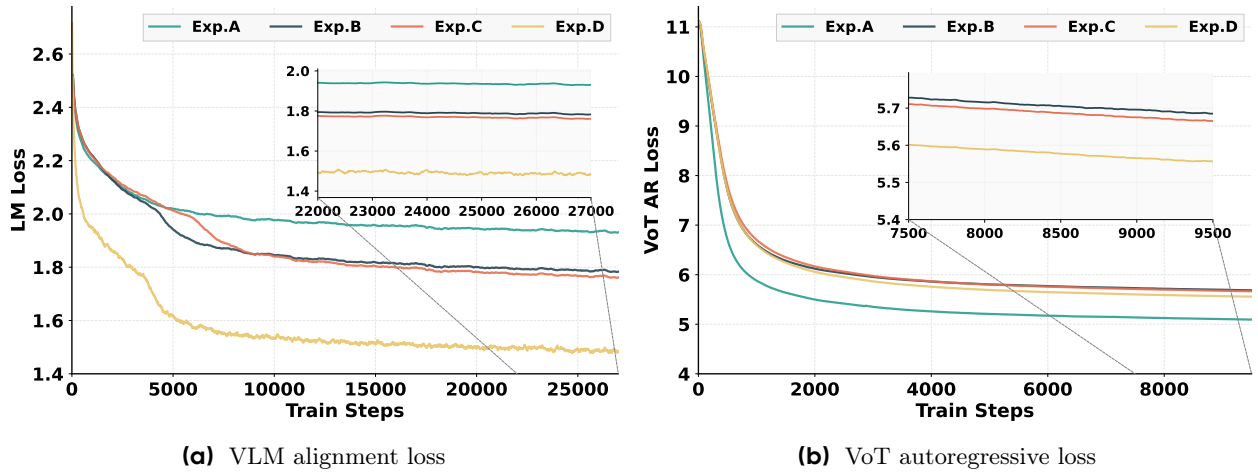


Figure 6 Training curves for the VLM alignment loss (left) and VoT autoregressive loss (right).

Visual Comparison. To qualitatively assess model capabilities in challenging scenarios, we select representative cases involving complex layout control and counterfactual generation, as shown in Fig. 5. We compare our results with GPT-Image-1. Our model (left) accurately renders precise attributes (e.g., facial expressions) and surreal scenes (e.g., “a horse riding an astronaut”), whereas GPT-Image-1 often fails because of attribute leakage or reliance on training-distribution biases.

6.3 Ablation Study

To validate each component of VoT tokenizer training, we evaluate six configurations (A–F), as summarized in Table 2. Configurations A–D isolate the effects of training objectives, architectural adaptation, and teacher scaling; E and F evaluate the full VoT model before and after post-training.

Table 2 Ablation study of tokenizer configurations. We progressively integrate reconstruction loss, trainable adapters, and teacher scaling. **Enc/Dec:** trainable encoder/decoder layers. **Align/Rec:** VLM alignment/reconstruction loss. All tokenizer variants use $\mathcal{L}_{VQ}$. $\dagger$ denotes the final model after supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF).

Exp.	Teacher VLM	Enc. / Dec. Depth	$\mathcal{L}_{align}$	$\mathcal{L}_{rec}$	AutoEval
Mogao (baseline)	–	– / –	–	–	53.12
A	Qwen2-VL 2B	– / –	✓	–	58.67
B	Qwen2-VL 2B	– / 2	✓	✓	60.98
C	Qwen2-VL 2B	1 / 2	✓	✓	64.64
D	Qwen2.5-VL 7B	1 / 2	✓	✓	69.32
E (VoT)	Qwen2.5-VL 7B	1 / 6	✓	✓	70.37
F (VoT $\dagger$)	Qwen2.5-VL 7B	1 / 6	✓	✓	76.22

Benefits of Reconstruction and Adapters. In Fig. 6a, LM loss drops from ~ 1.93 (Exp.A) to ~ 1.79 with reconstruction (Exp.B) and ~ 1.76 with a 1-layer adapter (Exp.C). These components stabilize token learning and reduce the mismatch with frozen ViT features; AutoEval correspondingly rises from $58.67 \rightarrow 60.98 \rightarrow 64.64$.

VoT AR Loss vs. Tokenizer Quality. Fig. 6b shows the VoT autoregressive loss in Stage 1 (Sec. 5). Exp.A achieves the lowest final autoregressive loss (~ 5.08) yet has the highest tokenizer LM loss in Fig. 6a (~ 1.93). AutoEval further confirms that predictability is not synonymous with utility: despite its lower autoregressive loss, Exp.A obtains only 58.67, substantially below Exp.B and Exp.C. Thus, reconstruction and the adapter can slightly increase autoregressive difficulty by preserving more information while improving semantic alignment and downstream utility.

Effect of Teacher VLM Scale. Comparing Exps. A–C (Qwen2-VL 2B teacher) with Exp.D (Qwen2.5-VL 7B teacher), Fig. 6a shows the LM loss dropping from ~ 1.76 – 1.93 to ~ 1.48 , while AutoEval increases from 64.64 to 69.32. Although reconstruction can make autoregressive prediction harder by preserving more information (Fig. 6b), Exp.D still achieves the lowest VoT autoregressive loss among the reconstruction-based settings. This result suggests that a stronger teacher provides cleaner semantic supervision and induces a more structured, autoregressive-friendly visual lexicon.

Benefits of VoT and Post-Training. We next compare the original *Mogao* baseline with our VoT model. Exp.E is the model after training with VoT tokens, making it the most direct comparison to *Mogao* under the same T2I training data. As shown in Table 2, VoT tokens yield a clear improvement in AutoEval ($53.12 \rightarrow 70.37$). Exp.F further improves the score after SFT and RLHF, indicating that post-training is complementary to VoT and translates better planning into higher final generation quality.

7 Conclusion

In this work, we introduced *Vision-of-Thought (VoT)*, a unified framework that bridges semantic understanding and visual synthesis by inserting a discrete visual planning layer between VLMs and diffusion decoders. By leveraging the VLM to generate structured VoT tokens as explicit plans before rendering, our approach successfully decouples high-level reasoning from pixel-level generation, significantly enhancing cross-modal alignment and interpretability. We hope VoT serves as a foundational step towards next-generation multimodal agents capable of coherent visual thinking and creation.

References

- [1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2023. URL <https://cdn.openai.com/papers/dall-e-3.pdf>.
- [2] Huiwen Chang, Han Zhang, Jarred Barber, Ajinkya Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.
- [3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.
- [4] Xiaokang Chen, Chengyue Wu, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- [5] Yao Chen, Siyuan Qian, Yuchen Chu, Zheng Liu, Kaipeng Zhang, Qifeng Chen, Wenping Wang, and Hao Li. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.
- [6] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuhan Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *AAAI*, volume 38, pages 1254–1262, 2024.
- [7] 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.
- [8] 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*, pages 12873–12883, 2021.
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, pages 12606–12633, 2024.
- [10] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2023.
- [11] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Planting a SEED of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.
- [12] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arxiv:2404.14396*, 2024.
- [13] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.
- [14] Yang Jin, Zhibin Xu, Ying Chen, , et al. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.
- [15] Black Forest Labs. Flux, 2024. URL <https://github.com/black-forest-labs/flux>.
- [16] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.
- [17] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arxiv:2402.08268*, 2024.
- [18] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations*, 2023.

- [19] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.
- [20] OpenAI. Dall-e 3 system card, 2023. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf.
- [21] OpenAI. Gpt-4o system card, 2024. URL <https://openai.com/index/gpt-4o-system-card/>.
- [22] 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.
- [23] 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. In *ICLR*, 2024.
- [24] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.
- [25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arxiv:2204.06125*, 2022.
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.
- [27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- [28] Quanhan Sun, Yuying Ge, Xuyang Lu, Haodong Chen, Peng Wang, Xufeng Li, Kaipeng Chen, Jingyi Zhang, Rui Zhang, Aojun Li, et al. Generative pretraining in multimodality. *arXiv preprint arXiv:2312.13286*, 2023.
- [29] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [30] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. ILLUME: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024.
- [31] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. EMU-3: Next-token prediction is all you need. *arXiv preprint arxiv:2409.18869*, 2024.
- [32] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-O: One single transformer to unify multimodal understanding and generation. *arXiv preprint arxiv:2408.12528*, 2024.
- [33] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xGEN-MM (BLIP-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.
- [34] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A Ross, Patrick Esser, Yonatan Bisk, Ming-Hsuan Yang, et al. Language model beats diffusion – tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2024.
- [35] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.
- [36] 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.

Appendix

A Training Configurations

Table 3 Training compute budget measured in GPU-hours on high-performance GPUs.

Stage	GPU-hours	Share (%)
Pretraining (tokenizer and unified model)	360 000	75.0
Continued training (CT)	50 000	10.4
Supervised fine-tuning (SFT)	64 000	13.3
Reinforcement learning (RLHF)	6000	1.3
Total	480000	100.0

B Limitations and Future Work

While VoT effectively bridges multimodal understanding and generation, its MoT architecture uses a frozen VLM backbone with parallel experts, limiting the system to unidirectional reasoning-to-generation transfer. This design preserves the VLM’s existing capabilities, but future work will explore bidirectional optimization strategies that jointly improve understanding and generation without catastrophic forgetting.

C Additional Text-to-Image Generation Results



Figure 7 Additional text-to-image generation results.