MENTOR: Efficient Autoregressive Image Generation with Balanced Multimodal Control

Anonymous ACL submission

Abstract

Recent text-to-image models achieve impressive visual quality but still face challenges in precise controllability, balancing multimodal inputs, and high training cost for multimodal image generation. To address these limitations, we propose **MENTOR**, an autoregressive (AR) framework with a two-stage training paradigm for controllable multimodal image generation: (1) a *multimodal alignment stage* that establishes robust pixel and semantic-level alignment between inputs and generated tokens, followed by (2) a *multimodal instruction tuning stage* that balances the model's integration of multimodal inputs and enhances generation controllability. Extensive experiments on DreamBench++ and DreamBench demonstrate that, despite modest model size and training resources, MENTOR achieves a strong balance between textual and visual guidance for controllable image generation, delivering competitive performance at significantly lower computational cost compared to leading baselines. Moreover, our approach attains superior image reconstruction fidelity, broad adaptability across different tasks, and training efficiency.

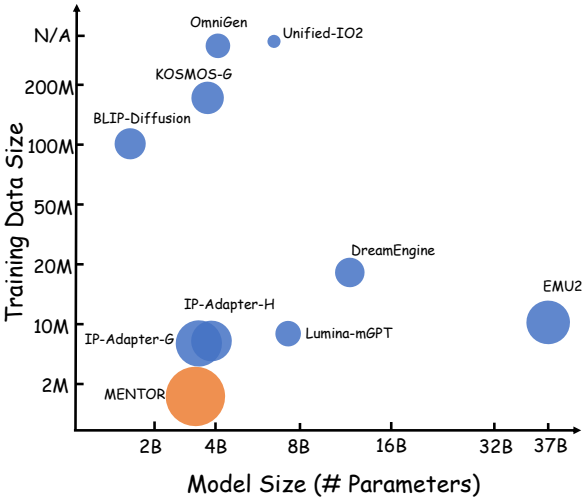

Figure 1: CP·PF score (circle size) of **MENTOR** and **other baselines** on DreamBench++. Model in lower left achieves the best efficiency.

1 Introduction

Recent progress in generative models has revolutionized text-to-image (T2I) generation (Ho et al., 2020; Rombach et al., 2022b; Podell et al., 2023). However, real-world applications often require more than text-only prompts. To achieve *controllable* image generation, e.g., fine-grained control over generated images, models need to seamlessly integrate multi-modal inputs, such as a reference image together with a detailed text prompt. This poses significant challenges for existing models that are predominantly focused on T2I generation. To address this, researchers integrate Large Multimodal Models (LMMs) with generative models (Pan et al., 2024; Sun et al., 2024c; Xiao et al., 2024a; Zhuang et al., 2025) to better handle multimodal inputs. Although effective for tasks like interleaved image-text generation (Sun et al., 2024c) and multimodal in-context learning (Pan et al., 2024; Ge et al., 2024), these approaches still face three challenges when scaling to *complex multimodal control*, especially under limited resources:

First, stochastic sampling in diffusion processes makes *precise and deterministic* control difficult, which is essential for high-fidelity tasks (e.g., faithful reconstruction) (Wang et al., 2025). **Second,** balancing guidance across modalities remains challenging. Existing methods exhibit modality imbalance, overemphasize one modality while neglecting the other (Han et al., 2024b). This phenomenon appears in both diffusion and autoregressive (AR) paradigms—for instance, IP-Adapter (Ye et al., 2023) and Lumina-mGPT (Zhuo et al., 2024), conditioned on text and image features, tend to favor image guidance. Such imbalance may stem from modality gaps, architectural limitations (Zhao et al., 2023; Cao et al., 2025; Ye et al., 2023), or suboptimal training schemes (Pan et al., 2024; Han

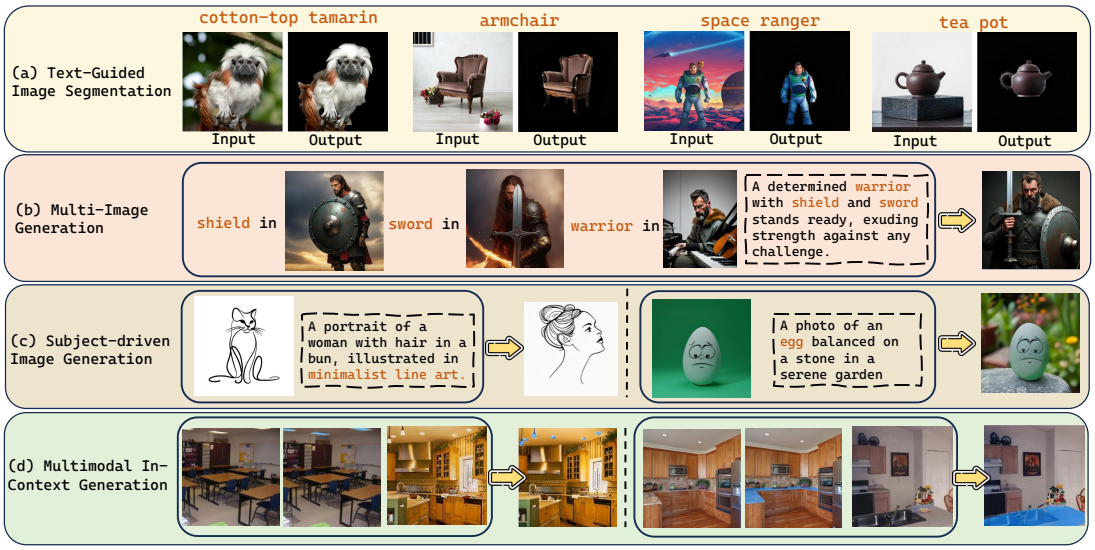

Figure 2: Qualitative examples of different tasks built on MENTOR after simply fine-tuning.

et al., 2024b). **Third,** existing methods rely on auxiliary alignment components, such as learned adapters (Pan et al., 2024), regression heads (Sun et al., 2023, 2024c) or specialized embeddings (Ge et al., 2023)), and demand large-scale training (Sun et al., 2024c; Pan et al., 2024; Ge et al., 2024), leading to significant computational costs. These observations motivate a question: *Is it possible to design an efficient framework for controllable multimodal image generation under limited resources?*

To address these limitations, we propose **MENTOR**, an efficient autoregressive (AR) framework for controllable multimodal image generation. Unlike diffusion models that demand complex cross-attention layers for multimodal conditioning and extensive training resources (Sun et al., 2024c; Pan et al., 2024; Li et al., 2023a), MENTOR employs a unified transformer that directly aligns multimodal inputs with output tokens, simplifying the architecture, removing auxiliary alignment modules, and substantially reducing training costs. Our framework uses a multimodal encoder to project inputs into a unified representation, which a transformer decoder then uses to deterministically generate image tokens. To ensure effective and balanced modality integration (Han et al., 2024b), we further adopt a two-stage training paradigm: (1) a *multimodal alignment stage* that builds robust pixel- and semantic-level alignment between inputs and generated tokens, followed by (2) a *multimodal instruction tuning stage* that balances the modality fusion and enhances generation controllability.

Notably, despite its simplicity and usage of sub-optimal checkpoints, MENTOR achieves competitive performance on various benchmarks including Dreambench (Ruiz et al., 2023) and Dreambench++ (Peng et al., 2025) with over 10× less resources than leading baselines. It surpasses resource-intensive models equipped with powerful generators such as SDXL (Podell et al., 2023), SD3 (Esser et al., 2024) and Infinity (Han et al., 2024a). Controlled experiments show that MENTOR offers a favorable trade-off between efficiency and multimodal balance. Further analyses highlight its adaptability across reinforcement learning and diverse multimodal tasks, offering a practical framework for controllable multimodal generation.

Overall, our contributions are as follows: (1) An autoregressive framework for efficient controllable multimodal image generation; (2) A two-stage training strategy that enables robust alignment and balanced modality integration with substantially reduced computational cost; (3) Experiments demonstrating the superior efficiency, controllability, and fidelity of MENTOR as a compelling framework for controllable multimodal generation.

2 Method

2.1 Model Design

As illustrated in Figure 3, MENTOR architecture comprises two core components: a multimodal encoder and an autoregressive generation decoder. These components are designed to unify multimodal inputs into a shared embedding and generate image tokens sequentially conditioned on the unified embedding, respectively.

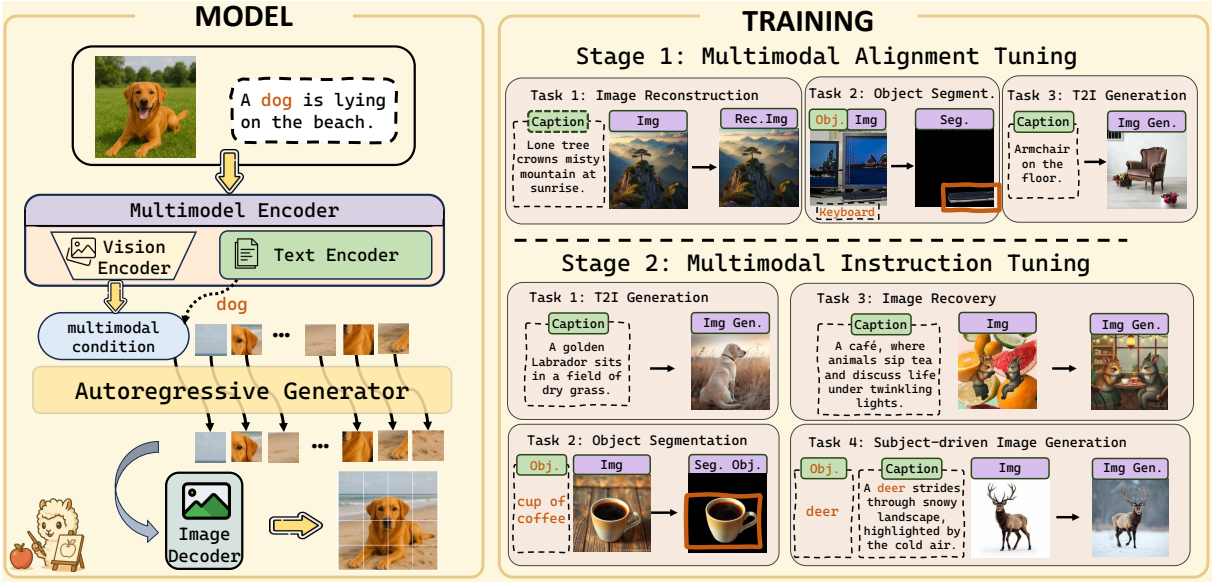

Figure 3: Overview of MENTOR. **Left** panel illustrates model structure, where visual and textual inputs are encoded into a unified latent to guide autoregressive image generation. **Right** panel highlights two-stage training paradigm: (1) **Multimodal Alignment Tuning**, enabling pixel and semantic-level alignment between inputs and output tokens; and (2) **Multimodal Instruction Tuning**, compels model to effectively balance influence of different modalities.

Multimodal Encoder The multimodal encoder integrates multimodal inputs from frozen pre-trained vision (ϕ_V) and language (ϕ_L) encoders into a shared latent space. This module projects visual features from ϕ_V into ϕ_L's embedding space using a lightweight connector module (ψ), yielding a unified multimodal representation $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_M)$, where $\mathbf{h}_j \in \mathbb{R}^d$. In practice, we adopt a multi-layer perceptron layer as the connector module that directly projects visual tokens to maintain detail information.

Autoregressive Decoder A transformer-based autoregressive decoder generates a image token sequence $\mathbf{y} = (y_1, \ldots, y_L)$ conditioned on the prefix $\mathbf{H}$ generated by the multimodal encoder and previously generated tokens $y_{<i}$. It operates in a shared embedding space with the encoder's output and shares the same vocabulary as the VQGAN (Esser et al., 2020) that is used for image tokenization. The generated token sequences are subsequently decoded into images using the VQGAN decoder. This unified autoregressive structure facilitates unified training via next-token prediction.

2.2 Two-Stage Training Paradigm

Effectively aligning different modalities and balancing their influence are crucial challenges for multimodal image generation. Our two-stage training paradigm directly addresses these issues, moving beyond initial coarse alignment to foster robust understanding and balanced integration of diverse inputs, as illustrated in Figure 3.

Stage 1: Multimodal Alignment Tuning. Although the multimodal encoder can encode multimodal inputs for the generator, we observe that the model tends to interpret visual information semantically like text captions, while neglecting visual details. To explicitly strengthen both pixel- and semantic-level alignment and promote richer visual utilization, we introduce three complementary tasks: (1) ***image reconstruction***, where the model reconstructs an input image (with or without corresponding caption) to enhance pixel-level fidelity; (2) ***object segmentation***, where the model generates an object-specific segmentation mask given an image and target label, enforcing attention to fine-grained visual details and spatial structures associated with semantic concepts; and (3) ***text-to-image (T2I) generation***, using image–caption pairs to preserve and reinforce generative capability. Notably, combining segmentation with reconstruction prevents trivial copy behavior (i.e., simply replicating input images) by encouraging semantically meaningful and spatially precise outputs. This complementary effect is further analyzed in § 3.3.

Stage 2: Multimodal Instruction Tuning. Building upon the alignment and visual fidelity established in Stage 1, this stage aims to equip the model with robust instruction following and cross-modal reasoning abilities. To encourage the

model to integrate different modalities in a balanced manner, we adopt a multimodal instruction tuning strategy using a curated mixture of tasks. We retain the *text-to-image (T2I) generation* and *object segmentation* tasks from Stage 1, preserving their formulations to maintain foundational skills and stabilize training. In addition, two new tasks are introduced to enhance instruction adherence and balanced modality integration. (1) ***image recovery*** introduces synthetic distortions—such as rotation, resizing, and compositing segmented objects onto random backgrounds—paired with original captions. The model are required to reconstruct the original image from the distorted input and caption, encouraging it to extract essential visual cues and leverage textual information to restore missing details. (2) ***subject-driven image generation*** conditions the model on a reference image, subject label, and textual instruction to generate new images. This task requires preserving the subject's visual identity while faithfully following textual directives, serving as a comprehensive end-to-end objective. Overall, this stage enables the model to integrate visual and textual information in a balanced manner, effectively mitigating over-reliance on a single modality and supporting precise multimodal generation, as further discussed in § 3.3.

2.3 Data Construction

To support our two-stage training paradigm, we construct a multimodal dataset of approximately **3 million samples**. The dataset integrates open-source resources, synthetic data, and automated annotations to ensure scalability and diversity. For **image reconstruction** and **T2I generation**, we collect image–text pairs from datasets such as CC12M (Changpinyo et al., 2021) and Midjourney-Niji (Emporium, 2024). To expand domain coverage (e.g., human subjects, artistic scenes), we further synthesize samples using T2I models like Flux.1 (BlackForestLabs, 2024), with prompts generated by advanced LLMs (OpenAI, 2024) to enhance semantic and visual diversity. For **segmentation** and **image recovery**, which require fine-grained object-level annotations, we design an automated pipeline that combines state-of-the-art LMMs (Wang et al., 2024a) with segmentation models(Kirillov et al., 2023). For **subject-driven image generation**, we leverage the OminiControl dataset (Tan et al., 2025), re-captioned using LMMs to accurately extract subject-relevant descriptions. Additionally, we reverse image pairs to effectively double the usable data. Data pipeline and formation are detailed in Appendix D.

3 Experiments

3.1 Experimental Setup

Implementation Details We initialized the multimodal encoder with CLIP-Large-Patch14 (Radford et al., 2021) and FlanT5-XL (Chung et al., 2022), which has 224×224 image receptive field, and the generator with LlamaGen-XL (Sun et al., 2024a). Training on 8 A100 GPUs about 1.5 days. More Details in Appendix B and Appendix I.2.

Benchmark & Metric. We evaluate MENTOR on **DreamBench** (Ruiz et al., 2023) and **DreamBench++** (Peng et al., 2025) benchmarks. **DreamBench** uses CLIP and DINO scores to assess image fidelity and prompt alignment. **DreamBench++** addresses limitations of DreamBench evaluation and evaluate on two axes: **Concept Preservation (CP)**, measuring the retention of the subject's visual identity, and **Prompt Following (PF)**, evaluating how accurately the image reflects the text prompt. A human evaluation study confirms high consistency between GPT-based metrics and human judgments. Further details are provided in Appendix E.1.

Baselines. We compare MENTOR with both fine-tuning and tuning-free methods. Fine-tuning baselines include **Textual Inversion** (Gal et al., 2022) and **DreamBooth** (Peng et al., 2025). Tuning-free baselines include diffusion-based models—**BLIP-Diffusion** (Li et al., 2023a), **Emu2** (Sun et al., 2024c), **IP-Adapter** (Ye et al., 2023), **OmniGen** (Xiao et al., 2024b), and **DreamEngine** (Chen et al., 2025b)—and AR models such as **Unified-IO 2** (Lu et al., 2023), **Lumina-mGPT** (Zhuo et al., 2024), and **VARGPT-v1.1** (Zhuang et al., 2025).

3.2 Main Results

Table 1 comprehensively evaluates our proposed autoregressive (AR) framework on the DreamBench++ benchmark, comparing it with diffusion-based and autoregressive-based baselines. MENTOR demonstrates highly competitive performance, particularly in achieving a strong balance between the guidance of both input modalities. Notably, this is achieved despite utilizing significantly fewer training resources and suboptimal model components compared to the state-of-the-art baselines.

Overall Performance. MENTOR achieves a strong balance between concept fidelity and prompt alignment, resulting in the high **CP·PF** score.

Table 1: Comparison on DreamBench++. Models are ranked by **CP·PF**, indicating balanced overall multimodal image generation performance. **CP/PF** ratio reflects overfitting issue toward certain modality. "*" denotes model trained **from scratch**; others are adapted from pre-trained T2I models.

Method	T2I Model	Train Data	Model Size	Concept Preservation (CP)					Prompt Following (PF)				CP·PF	CP/PF
				Animal	Human	Object	Style	Overall	Photo.	Style.	Imag.	Overall		
Finetuned on Test Set														
Textual Inv.	SD v1.5	-	860M	0.50	0.35	0.30	0.36	0.38	0.68	0.70	0.44	0.63	0.24	0.60
DreamBooth	SD v1.5	-	860M	0.64	0.19	0.49	0.48	0.49	0.79	0.78	0.51	0.72	0.36	0.68
DreamBooth-L	SDXL v1.0	-	2.60B	0.75	0.31	0.54	0.71	0.60	0.90	0.90	0.75	0.87	0.52	0.69
Test-Time Tuning-Free Methods														
VARGPT-v1.1	Infinity (VAR)	8.3M	9.00B	0.25	0.13	0.14	0.27	0.19	0.31	0.48	0.33	0.37	0.07	0.51
Unified-IO2*	Unified-IO2	8.5B	7.00B	0.77	0.80	0.64	0.82	0.72	0.24	0.18	0.11	0.19	0.14	3.74
Lumina-mGPT	Chameleon	10M	7.00B	0.95	0.97	0.89	0.85	0.91	0.31	0.25	0.15	0.25	0.23	3.63
OmniGen*	OmniGen	700M	3.80B	0.39	0.35	0.25	0.22	0.30	0.70	0.71	0.67	0.70	0.21	0.43
DreamEngine	SD3.5	21M	10.50B	0.76	0.72	0.61	0.73	0.68	0.44	0.37	0.25	0.37	0.26	1.84
BLIP-Diffusion	SD v1.5	130M	1.56B	0.67	0.56	0.47	0.51	0.55	0.58	0.52	0.30	0.50	0.27	1.11
Kosmos-G	SD v1.5	200M	3.00B	0.62	0.64	0.46	0.56	0.55	0.48	0.62	0.39	0.51	0.28	1.07
IP-A-Plus ViT-H	SDXL v1.0	10M	3.00B	0.90	0.84	0.76	0.91	0.84	0.50	0.39	0.28	0.41	0.35	2.02
Emu2	SDXL v1.0	16M	37.00B	0.67	0.55	0.45	0.44	0.53	0.73	0.73	0.56	0.69	0.36	0.76
IP-A ViT-G	SDXL v1.0	10M	2.50B	0.67	0.56	0.50	0.75	0.59	0.74	0.63	0.45	0.64	0.38	0.92
MENTOR	LlamaGen	3M	2.31B	0.65	0.36	0.57	0.47	0.56	0.86	0.85	0.80	0.84	**0.47**	0.66

Quantitative examples can be found in Figure 6. It rivals fine-tuned methods such as DreamBooth-LoRA while significantly outperforming test-time tuning-free baselines including OmniGen and DreamEngine. A key strength of MENTOR lies in its ability to harmoniously integrate multimodal inputs. Several strong baselines, including autoregressive models like Lumina-mGPT and Unified-IO2, achieve high Concept Preservation (CP) but extremely low Prompt Following (PF), leading to high **CP/PF** ratios and indicating over-reliance on visual references while neglecting textual guidance. Even for models like VARGPT-v1.1, which has the advanced VAR generator, despite producing visually appealing results, fail to consistently follow multimodal instructions. In contrast, MENTOR delivers low CP/PF scores compared to most baselines, demonstrating effective and controlled integration of both visual and textual guidance.

Training Efficiency. A notable advantage of MENTOR lies in its training efficiency. It is trained on only 3 million image-text pairs across two stages, substantially less than leading baselines, such as Emu2 (16M), Kosmos-G (200M), and DreamEngine (21M). Beyond the reduced data requirements, the training process is highly resource-efficient: the entire training process completes in 1.5 days with 8 GPUs. This contrasts sharply with other baselines, such as Kosmos-G, which necessitates 256 GPUs over three days. Despite this dramatically reduced computational and data budgets, MENTOR achieves competitive performance with balanced performance, highlighting its efficiency and effectiveness. Furthermore, MENTOR remains highly competitive in size compared to larger counterparts, highlighting our effectiveness.

Discussion and Connection to Methodology. The strong performance of MENTOR —particularly its balanced multimodal generation and training efficiency—stems from its **autoregressive nature** and **two-stage training paradigm**. The **autoregressive design**, which generates image tokens sequentially conditioned on a unified multimodal prefix, enables fine-grained, token-level alignment between multimodal inputs and outputs. This direct alignment improves controllability and ensures generated results accurately follow both text and visual guidance. Notably, **two-stage training paradigm** promotes balanced multimodal control, mitigating the dominance of any single modality and yielding significantly higher CP·PF scores compared to baselines. Notably, MENTOR achieves strong results despite using **relatively suboptimal components**. While baselines rely on advanced models such as Qwen-2.5 and SD3, our implementation employs Flan-T5 as the encoder and LlamaGen as the generator—both substantially weaker than their counterparts, as shown in Table 6. This highlights that the improvements are stem from methodological design rather than model or data scale.

3.3 Ablation Study

We conduct an ablation study on DreamBench++ and DreamBench focusing on two central questions: (1) How critical is Stage 1 for robust multimodal alignment? and (2) What role does each

Table 2: Ablation results on DreamBench++ and DreamBench.

Method	DreamBench++			DreamBench		
	CP	PF	CP·PF	DINOv1	CLIP-I	CLIP-T
w/o Obj. Seg. in Stage 1	0.252 ± 0.004	0.479 ± 0.005	0.121	56.113 ± 0.082	74.384 ± 0.071	23.965 ± 0.038
w/o Stage 1 Alignment	0.179 ± 0.002	0.673 ± 0.012	0.120	33.523 ± 0.111	67.705 ± 0.101	28.263 ± 0.155
w/o Image Recovery	0.661 ± 0.007	0.284 ± 0.004	0.188	74.471 ± 0.321	81.280 ± 0.094	24.210 ± 0.022
w/o Object Segmentation	0.412 ± 0.002	0.918 ± 0.003	0.378	57.221 ± 0.119	76.269 ± 0.084	31.078 ± 0.050
w/o Multimodal T2I Task	0.407 ± 0.004	0.910 ± 0.004	0.370	58.880 ± 0.143	76.529 ± 0.102	30.483 ± 0.002
MENTOR	0.555 ± 0.006	0.839 ± 0.002	**0.466**	70.853 ± 0.327	80.911 ± 0.053	29.071 ± 0.080

training task play in shaping model's multimodal generation behavior? Following prior work (Peng et al., 2025), we report **CP·PF** as a primary measure of multimodal image generation ability.

Importance of Stage 1: Foundational Multimodal Alignment. As shown in Table 2, removing Stage 1 leads to severe performance drop, underscoring its foundational role. On DreamBench++, the CP score drops from 0.555 to 0.179, indicating a major loss in visual identity preservation, with PF also significantly reduced. Similar trends appear on DreamBench. Ablating only **object segmentation task** in Stage 1 (*w/o Obj. Seg. in Stage 1*) also hampers model performance. While the remaining image reconstruction task supports pixel-level alignment, allowing for reconstruction of input images, it inadvertently leads the model to exhibit a copy-paste behavior, failing to capture semantic and visual information of input images. It confirms that reconstruction alone is insufficient for robust multimodal alignment. Overall, these results highlight that without Stage 1, the model struggles to ground visual concepts from images, severely impairing its visual preservation ability.

Contributions of Different Training tasks in Stage 2. The distinct contributions of different Stage 2 tasks are also evident in Table 2. Excluding the *Image Recovery* task leads to a sharp imbalance: while visual preservation metrics (CP, DINOv1, CLIP-I) show a notable increase, instruction following ability (PF and CLIP-T score) critically drops, showing an overfitting to visual features. This underscores that image recovery acts as a critical regularizer, encouraging the model to reconstruct incomplete visual contexts guided by text prompt, thereby fostering a balanced use of different modalities. Conversely, ablating either the *Object Segmentation* or the *Subject-Driven Image generation* significantly degrades visual preservation ability, as these tasks prompt the model to utilize the visual features of the input image effectively to generate images. These results demonstrate that image recovery ensures cross-modal balance, while object segmentation and subject-driven generation enhance the model's ability to extract and utilize detailed visual information for image generation. Additional task weight sweeping experiments in Appendix I.1 further confirm the CP·PF trade-off controlled by the task mixture.

3.4 Analysis

Efficiency and Effectiveness: AR vs. Diffusion. To evaluate the efficiency of our AR framework against diffusion-based methods, we conducted a controlled comparison with **Kosmos-G** (Pan et al., 2024), a representative LMM-augmented diffusion model.Both models were trained from **similar initializations** on **same training data** to ensure a fair comparison. Shown in Table 3, despite Kosmos-G employing a superior SD1.5 generator and Kosmos-1 encoder, MENTOR, still shown significantly better performance with limited training data, highlighting the effectiveness of our framework.

Comparison of Architecture Variants. We compare architectural variants of MENTOR that differ in how visual features are connected to the generator. Since the multimodal encoder produces hundreds of visual tokens per image, the inflated context length poses significant computational challenges, particularly in multi-image settings. Token compression alleviates this issue but may compromise fine-grained visual fidelity. To navigate this trade-off, as shown in Table 2, we evaluate a query-based connector that leverages a Query-Former (Li et al., 2023b) to compress long visual sequences into fixed-size representations, guided by textual queries (Li et al., 2023a) highlighting important concepts in images. The results reveals that while the query-based approach substantially reduces computational cost, it struggles to preserve fine-grained visual details crucial for generative fidelity, even guided by textual queries. Consequently, the MLP-based connector(MENTOR) achieves higher fidelity, especially for human and object regions,

Table 3: Controllable experiments between MENTOR and Kosmos-G in DreamBench++ benchmark.

Method	Concept Preservation (CP)					Prompt Following (PF)				CP·PF
	Animal	Human	Object	Style	Overall	Photo.	Style.	Imag.	Overall	
Kosmos-G	0.17	0.08	0.14	0.18	0.15	0.72	0.71	0.68	0.71	0.11
MENTOR	0.65	0.36	0.57	0.47	**0.55**	0.86	0.85	0.80	**0.84**	**0.47**

Table 4: Image reconstruction performance.

Method	COCO (↓)	JourneyDB (↓)
SeedTokenizer	0.5102	0.5291
SEED-X	0.4317	0.4352
EMU2-Gen	0.3828	0.2869
DreamEngine	0.2065	0.2052
MENTOR	**0.1008**	**0.0867**

whereas the query-based variant offers a favorable trade-off between efficiency and performance. Despite these differences, both variants exhibit competitive performance compared to other baselines.

Effect of Multi-Image Training. To assess the benefits of richer visual context, we further trained the model using a mix of Stage 2 data and additional multi-subject task(reconstructing images based on segmented objects and image caption) generated via our data construction pipeline. As shown in Table 5, *w. MultiImage Training* achieves a higher CP·PF score (0.49), improving CP to 0.60 while maintaining a strong PF score. This emphasizes the advantage of enhanced visual context in training, prompting the model to efficiently handle and integrate information from multiple visual inputs, thereby improving its ability to preserve visual details in complex multimodal scenarios.

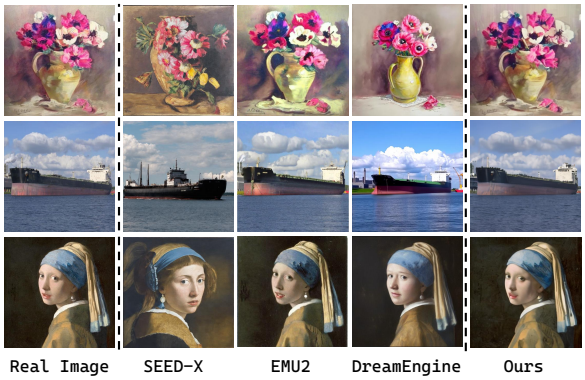

Figure 4: Qualitative study on Image Reconstruction.

Image Reconstruction Fidelity. To quantify visual detail preservation in our framework, we evaluate MENTOR on the Image Reconstruction Benchmark (Chen et al., 2025b), which measures similarity between input and reconstructed images. After fine-tuning on reconstruction task for 1,000 steps, we compare the generated outputs with their originals, following previous work (Chen et al., 2025b). As shown in Table 4, MENTOR outperforms strong baselines with comparable architectures—SeedTokenizer (Ge et al., 2023), EMU2 (Sun et al., 2024c), SeedX (Ge et al., 2024), and DreamEngine (Chen et al., 2025b)—all of which couple LMMs with diffusion backbones. MENTOR achieves best reconstruction quality even with a 224×224 receptive field, while others varied from 384x384 to 512x512.

Versatility Across Different Multimodal Tasks. To explore broader applicabilities of our framework, we evaluate its adaptability across diverse tasks, including image segmentation, multi-image generation and multimodal in-context image generation. This was achieved with brief fine-tuning on relevant datasets, as detailed in Appendix G. Qualitative results in Figure 2, Figure 11, Figure 14 and Figure 15 show that MENTOR produces coherent outputs that adhere to provided constraints without any architectural modifications. While achieving performance in each specific domain would necessitate more specialized training and potentially more powerful components, these initial results underscore our framework's versatility and its potential as an effective foundation for multimodal image generation.

Effect of Reinforcement Learning on Multimodal Generation. Given the autoregressive nature of our framework, MENTOR naturally supports token-level reinforcement learning (RL), enabling direct optimization of generation behaviors. We apply the GRPO (Shao et al., 2024) following the AR-GRPO (Yuan et al., 2025) to explore potential of RL in enhancing multimodal controllable generation. Specifically, the model is fine-tuned with reward signals from combined reward models, encouraging outputs that better align with multimodal guidance while preserving visual coherence. As shown in Table 5, *w.GRPO* achieves the high-

Table 5: Analysis on architecture design, multi-image training and reinforcement learning

Method	DreamBench++			DreamBench		
	CP	PF	CP·PF	DINOv1	CLIP-I	CLIP-T
MENTOR	0.555 ± 0.006	0.839 ± 0.002	0.466	70.853 ± 0.327	80.911 ± 0.053	29.071 ± 0.080
w. Query-Variants	0.421 ± 0.002	0.882 ± 0.000	0.371	54.518 ± 0.317	76.306 ± 0.114	30.792 ± 0.040
w. Multi-image	0.586 ± 0.006	0.829 ± 0.005	0.486	72.487 ± 0.147	81.857 ± 0.152	28.545 ± 0.043
w. GRPO	0.609 ± 0.006	0.866 ± 0.004	**0.527**	77.176 ± 0.199	83.343 ± 0.287	29.810 ± 0.059

est CP·PF score (0.527), surpassing other variants. More details can be found in Appendix B.4.

4 Related Work

4.1 Image Generation with Complex Multimodal Control

Recent advances in diffusion models enable image generation conditioned on multimodal inputs such as canny edges (Zhang et al., 2023) or reference images (Zhao et al., 2024; Meng et al., 2022). DreamBooth (Ruiz et al., 2023) enables subject-specific fine-tuning but limits generalization. SuTI (Chen et al., 2023b) addresses it with scalable data and training. To enhance flexibility, recent work integrates LMMs with diffusion models (Koh et al., 2023; Sun et al., 2023; Dong et al., 2023; Li et al., 2024). Approaches like Kosmos-G (Pan et al., 2024), Emu-2 (Sun et al., 2024c), Seed-X (Ge et al., 2024), and DreamEngine (Chen et al., 2025b) explore more complex multimodal prompt and fine-grained multimodal control. Yet, balancing guidance from diverse modalities remains a core challenge (Han et al., 2024b; Ye et al., 2023; Mao et al., 2024). EMMA (Han et al., 2024b) employs a gated perceiver resampler for dynamic signal integration, while RealCustom++(Mao et al., 2024) disentangles subject identity and textual fidelity via cross-layer projectors. OmniControl (Tan et al., 2025) introduces a bias term into multimodal attention. Recently, large-scale systems such as Qwen-Image (Wu et al., 2025a), Begal (Deng et al., 2025), and Nano-Banana (Comanici et al., 2025) further demonstrate impressive multimodal generation capabilities through large-scale training on extensive multimodal corpora across thousands of GPUs. Nonetheless, these methods still demand substantial resources, and achieving efficient multimodal integration remains an open problem.

4.2 Autoregressive Multimodal Generation

Autoregressive models have driven progress in T2I generation, from DALL·E (Ramesh et al., 2021) to LlamaGen (Sun et al., 2024a) and GPT4O (OpenAI, 2024). Recent work extends it to multimodal settings: Models like Chameleon (Team, 2024a), LWM (Liu et al., 2024a), AnyGPT (Zhan et al., 2024), and EMU3 (Wang et al., 2024b) treat text and images as unified token sequences via early-fusion transformers, yet still emphasize text-to-image generation with limited support for multimodal conditioning. Janus (Wu et al., 2024a) decouples visual understanding and generation via distinct pathways but lacks support for multimodal image generation. MUSE-VL (Xie et al., 2025) and VILA-U (Wu et al., 2024c) align discrete visual tokens with text to improve perception, but remain oriented toward understanding tasks rather than image generation. Unified-IO2 (Lu et al., 2023) is trained autoregressively from scratch for both understanding and generation across modalities, while Lumina-mGPT (Zhuo et al., 2024) and VARGPT (Zhuang et al., 2025) enhances Chameleon and VAR (Han et al., 2024a) with supervised fine-tuning for broader multimodal tasks. Overall, while models like VILA-U (Wu et al., 2024c), EMU3 (Wang et al., 2024b), and Janus (Wu et al., 2024a) have advanced text-to-image generation, robust multimodal conditional image generation remains an open and underexplored challenge.

5 Conclusion

In this work, we introduced a controllable and efficient autoregressive framework for complex multimodal image generation, offering a compelling alternative to diffusion-based methods. By unifying multimodal inputs within an AR model and leveraging a two-stage training paradigm, our method achieves state-of-the-art performance on challenging benchmarks—despite a modest model size, suboptimal base component, and limited training resources. These results underscore the efficiency, scalability, and controllability of our method, establishing it as an efficient foundation for building versatile, fine-grained visual generation systems capable of handling complex multimodal prompts.

Limitations

Our work introduces a resource-efficient framework for multimodal-conditioned image generation, primarily aimed at exploring how to balance multimodal guidance for controllable generation rather than maximizing absolute generation quality. Accordingly, our evaluation focuses on verifying the effectiveness of the proposed training paradigm and architectural design under limited-resource settings, using models of comparable capacity and scale. However, the current performance of our approach is inherently constrained by the limitations of generative backbone models. Unlike recent unified or large-scale models trained with thousands of GPUs or on extensive datasets, our experiments are conducted on a smaller scale using publicly available resources. Consequently, MENTOR exhibits shortcomings in text-to-image generation, including spatial reasoning, object counting, fine-grained human rendering, and stylization. These limitations reflect the current gap between current SOTA diffusion and autoregressive architectures in terms of generation fidelity and domain generalization. Future work can readily extend this framework by replacing backbone modules and scaling data resources to further enhance image quality. Additionally, while our training data is sourced from publicly available datasets and our synthetic data pipeline includes NSFW safeguards, a comprehensive evaluation of safety, fairness, and potential misuse remains lacking. Future work should incorporate thorough assessments of model biases and unintended behaviors. Finally, while our framework demonstrates strong versatility across diverse multimodal tasks, achieving competitive performance in specific domains may require more specialized training and the integration of more powerful multimodal encoders and generators.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science*.

BlackForestLabs. 2024. Announcing black forest labs.

Bing Cao, Baoshuo Cai, Changqing Zhang, and Qinghua Hu. 2025. Dig2dig: Dig into diffusion information gains for image fusion. *Preprint*, arXiv:2503.18627.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568.

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and 1 others. 2023a. PixArt-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. 2025b. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. *Preprint*, arXiv:2502.20172.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. 2023b. Subject-driven text-to-image generation via apprenticeship learning. *Preprint*, arXiv:2304.00186.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv: 2309.11499*.

Caption Emporium. 2024. midjourney-niji-1m-llavanext. https://huggingface.co/datasets/CaptionEmporium/conceptual-captions-cc12m-llavanext.

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

Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878.

Mark Everingham and John Winn. 2011. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 8(5):2–5.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *Preprint*, arXiv:2208.01618.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2024. GenEval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. 2024a. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *Preprint*, arXiv:2412.04431.

Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. 2024b. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts. *Preprint*, arXiv:2406.09162.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *ICCV*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating images with multimodal language models. *NeurIPS*.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. 2022. Language-driven semantic segmentation. *Preprint*, arXiv:2201.03546.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. 2023a. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Preprint*, arXiv:2305.14720.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. 2024. UNIMO-G: Unified image generation through multimodal conditional diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6173–6188, Bangkok, Thailand. Association for Computational Linguistics.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with blockwise ringattention. *Preprint*, arXiv:2402.08268.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024b. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *Preprint*, arXiv:2312.17172.

Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. 2024. Realcustom++: Representing images as real-word for real-time customization. *Preprint*, arXiv:2408.09744.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. *Preprint*, arXiv:2108.01073.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 891–898.

OpenAI. 2024. hello-gpt-4o.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-g: Generating images in context with multimodal large language models. *Preprint*, arXiv:2310.02992.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2025. Dreambench++: A human-aligned benchmark for personalized image generation. *Preprint*, arXiv:2406.16855.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *Preprint*, arXiv:2102.12092.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *Preprint*, arXiv:2208.12242.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

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

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

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

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

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2025. Ominicontrol: Minimal and universal control for diffusion transformer. *Preprint*, arXiv:2411.15098.

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

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

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformer Reinforcement Learning.

Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Xiaoming Wei, and Enhua Wu. 2025. Image editing with diffusion models: A survey. *Preprint*, arXiv:2504.13226.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, and 6 others. 2024b. Emu3: Next-token prediction is all you need. *Preprint*, arXiv:2409.18869.

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

Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. 2024. Omniedit: Building image editing generalist models through specialist supervision. *arXiv preprint arXiv:2411.07199*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025a. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2024a. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *Preprint*, arXiv:2410.13848.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and 1 others. 2024b. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. 2025b. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 2024c. Vila-u: a unified foundation model integrating visual understanding and generation. *Preprint*, arXiv:2409.04429.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. 2024a. Omnigen: Unified image generation. *Preprint*, arXiv:2409.11340.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. 2024b. Omnigen: Unified image generation. *Preprint*, arXiv:2409.11340.

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

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. 2025. Muse-vl: Modeling unified vlm through semantic discrete encoding. *Preprint*, arXiv:2411.17762.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.

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

Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. 2025. Ar-grpo: Training autoregressive image generation models via reinforcement learning. *Preprint*, arXiv:2508.06924.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *ArXiv*, abs/2402.12226.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *ArXiv preprint*, abs/2309.07915.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2024. Ultraedit: Instruction-based fine-grained image editing at scale. *Preprint*, arXiv:2407.05282.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641.

Xianwei Zhuang, Yuxin Xie, Yufan Deng, Dongchao Yang, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. 2025. Vargpt-v1.1: Improve visual autoregressive large unified model via iterative instruction tuning and reinforcement learning. *Preprint*, arXiv:2504.02949.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, and 1 others. 2024. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. *arXiv preprint arXiv:2406.18583*.

A Appendix

This Appendix is organized as follows.

- In Appendix A.1, we describe the autoregressive training objective for our framework.
- In Appendix B, we provide the training details of MENTOR, including initialization (Appendix B.1), training procedures (Appendix B.2), multi-image training strategy (Appendix B.3), and RL training with GRPO (Appendix B.4).
- In Appendix C, we show quantitative evaluations of our method on text-to-image generation benchmarks.
- In Appendix D, we detail our data construction pipeline and the dataset details used across the two-stage training.
- In Appendix E, we elaborate on the experimental setup, including datasets and metrics (Appendix E.1), as well as detailed descriptions of baseline methods (Appendix E.2).
- In Appendix F, we present qualitative results that demonstrate the capabilities of MENTOR in various settings, such as image reconstruction (Appendix F.1), segmentation (Appendix F.2), multi-image generation (Appendix F.3), and in-context image generation (Appendix F.4).
- In Appendix G, we demonstrate the versatility of MENTOR across diverse multimodal generation tasks, including segmentation, subject-driven generation, and multimodal in-context learning.
- In Appendix H, we present the human evaluation study validating the reliability of GPT-based metrics.
- In Appendix I, we analyze CP–PF trade-offs via task weight and CFG sweeps.

A.1 Preliminary

Training Objective Our model employs *teacher forcing* to predict image tokens, conditioned on (i) previously generated tokens and (ii) multimodal context $\mathbf{h}$. Given the multimodal condition: $\mathbf{c}^{(0)} = \{\mathcal{I}, \mathcal{T}\}$ (visual and textual inputs), a multimodal encoder ϕ first encodes $\mathbf{c}^{(0)}$ and subsequently uses an MLP layer to project them into the space of the image decoder to form a unified representation $\mathbf{h}$:

$$\mathbf{H} = \text{MLP}(\phi(\mathbf{c}^{(0)})) = (\mathbf{h}_1, \ldots, \mathbf{h}_M) \in \mathbb{R}^{M \times d}, \mathbf{h}_j \in \mathbb{R}^d. \tag{1}$$

where M is the number of conditioning tokens, and d is the dimension of the latent embeddings. Then, the AR decoder θ, conditioned on $\mathbf{h}$, generates image sequence $\mathbf{y} = (y_1, \ldots, y_L)$ as follows:

$$\theta(\mathbf{y} \mid \mathbf{H}) = \prod_{i=1}^{L} \theta(y_i \mid y_{<i}, \mathbf{H}). \tag{2}$$

The training objective is to minimize the token-level cross-entropy loss by *teacher forcing* on data $\mathcal{D}$:

$$\mathcal{L}_{\text{CE}}(\theta, \phi) = -\mathbb{E}_{(\mathbf{y}, \mathbf{c}^{(0)}) \sim \mathcal{D}} \left[\sum_{i=1}^{L} \log \theta(y_i \mid y_{<i}, \mathbf{H}) \right]. \tag{3}$$

Classifier-free Guidance To enhance multimodal generation controllability, we apply Classifier-Free Guidance (CFG) (Sun et al., 2024a). During training, multimodal conditioning $\mathbf{H}$ is replaced by a learned unconditional embedding $\mathbf{H}_u$ with probability p. At inference time, token logits ℓ_g are recalculated by interpolating between the conditional logits ℓ_c (from $\mathbf{H}$) and unconditional logits ℓ_u (from $\mathbf{H}_u$), controlled by a scaling parameter λ: $\ell_g = \ell_u + (\ell_c - \ell_u) \times \lambda$.

B Training Details

B.1 Initialization Details

The multimodal encoder is initialized using the vision encoder from CLIP-Large-Patch14 (Radford et al., 2021), with an image receptive field of 224×224, and the FlanT5-XL encoder (Chung et al., 2022), with a context length of 512 tokens. This encoder converts each image into 256 tokens for use as context in the generator.

To implement the MLP-based projection, we train the MLP projector on the LLaVA-CC3M-Pretrain-595K dataset (Liu et al., 2023), following the alignment training setup used by LLaVA. Specifically, we freeze both the vision and text encoders (CLIP-Large-Patch14 and FlanT5-XL, respectively) and train only the MLP layers. The resulting pretrained MLP layers are then directly incorporated into the multimodal encoder of MENTOR.

The projector consists of a two-layer MLP with an intermediate dimension of 4,096, employing SiLU activation functions. The autoregressive generator is initialized from LlamaGen-XL (Sun et al.,

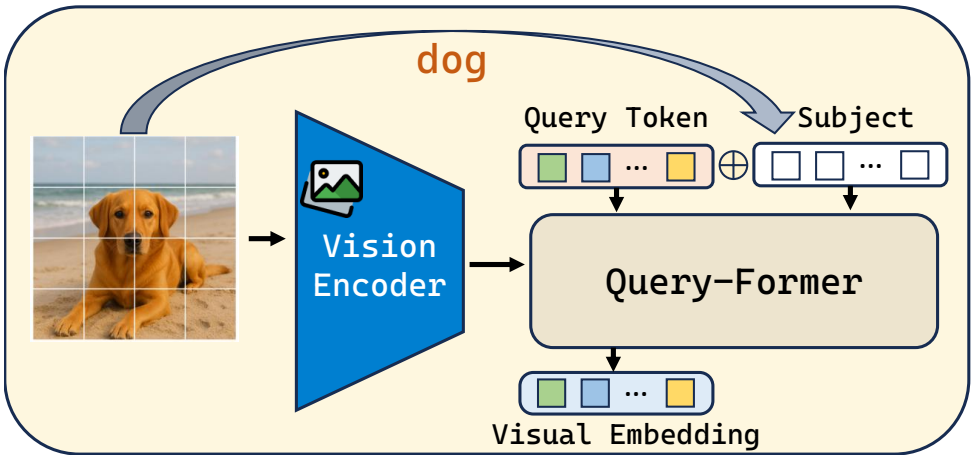

Figure 5: Overview of text-guided visual distillation using the Query-based variant of MENTOR.

2024a) with 775 million parameters. However, the original LlamaGen implementation contains a fundamental error in its 2D Rotary Positional Embedding (Lu et al., 2023; Fang et al., 2024) (ROPE) mechanism*, which leads to a loss of information in the query and key vectors during attention computation. To address this, we correct the ROPE implementation in our code and continue training the revised model on both the Midjourney dataset (Emporium, 2024) and the LAION-COCO dataset used in LlamaGen pretraining, effectively replicating the original pretraining conditions. This continued training enables the model to adapt to the corrected ROPE mechanism. The resulting model is then used to initialize our autoregressive generator.

B.2 Training Procedure

The model training comprises two distinct stages:

Stage 1: We freeze the multimodal encoder and train only the projector and generator modules for one epoch, using a global batch size of 128. Optimization employs the Adam optimizer with an initial learning rate of 5×10^{-4}, a linear warm-up over the initial 5

Stage 2: We fine-tune the entire model, excluding the vision encoder, for two epochs. The learning rate is reduced to 1×10^{-4}, with all other optimization settings remaining consistent with Stage 1. This phase primarily enhances cross-modal interactions and improves conditional image generation capabilities from combined visual and textual inputs.

Training is conducted across 8 NVIDIA A100 GPUs, each equipped with 80 GB memory, taking approximately 1.5 days. Specifically, Stage 1 training involves 2.48 million data points over a single epoch, completed in roughly 14 hours. Stage 2 training utilizes 1.3 million data points over two epochs, taking approximately 20 hours in total.

Ablation studies follow the same training schedule, with one epoch of training on Stage 1 data, followed by two epochs on Stage 2 data.

B.3 Multi-Image Training

In the **multi-image training** scenario, the context length of MENTOR is expanded to 1,280 tokens to accommodate up to 4 images per context. For the Query-based variant of MENTOR, token compression techniques enable processing up to 14 images per context with 512 context length.

We utilize 1.5 million multi-image samples, each comprising segmented sub-images accompanied by textual descriptions. The model is trained to reconstruct original images based on these segmented inputs and their corresponding captions. Training incorporates a mixture of Stage 2 data and multi-image samples for an additional epoch.

Qualitative assessments, presented in Figure 7, demonstrate that multi-image training significantly enhances the model's capability to preserve detailed visual information in complex multimodal contexts.

B.4 GRPO Fine-Tuning for Autoregressive Multimodal Generation

To explore reinforcement learning for controllable multimodal generation, we adopt the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) under the autoregressive (AR) training paradigm, following the AR-GRPO setup (Yuan

*https://github.com/FoundationVision/LlamaGen/issues/54

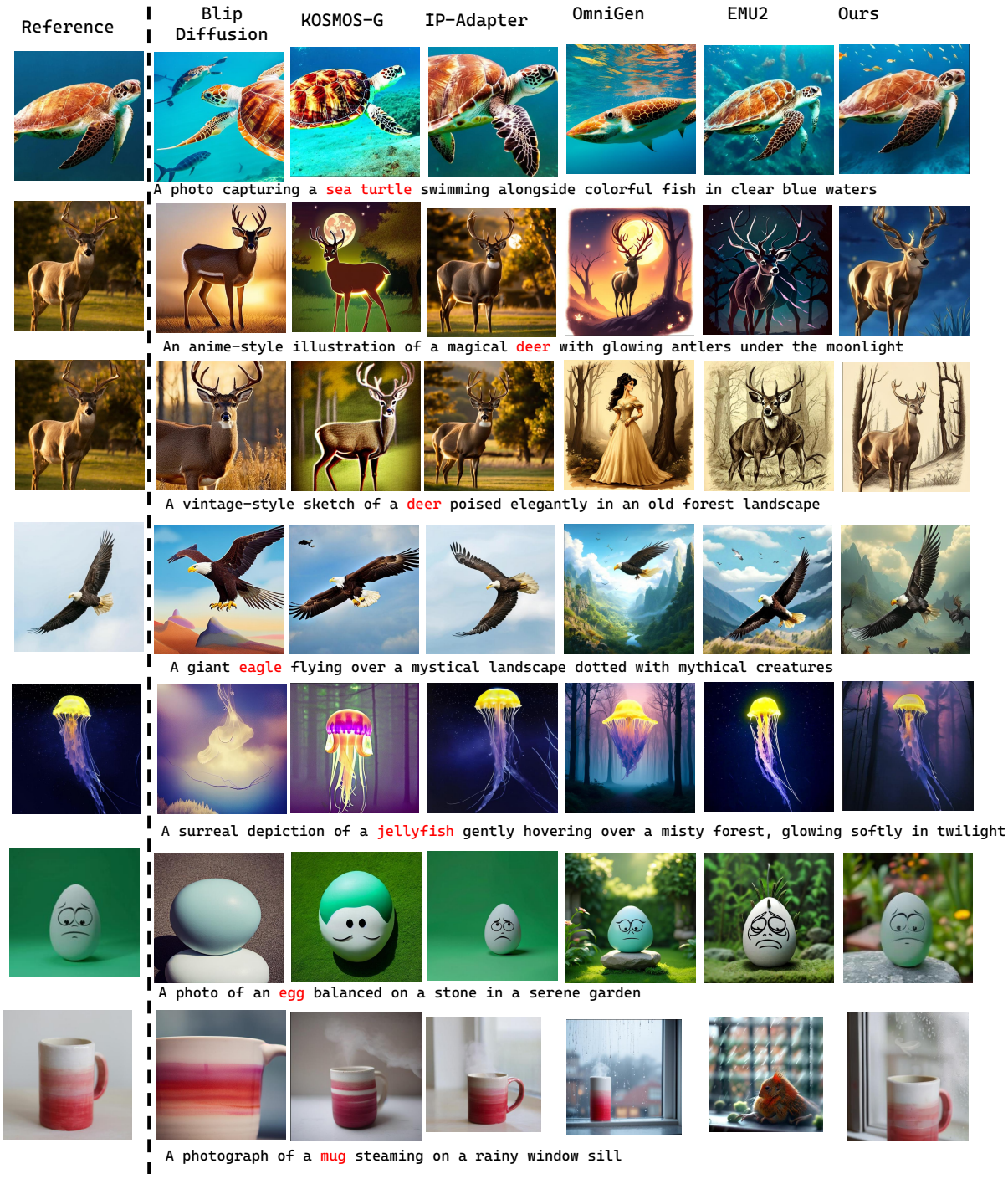

Figure 6: Qualitative examples of different methods compared to MENTOR on DreamBench++.

et al., 2025). Given the autoregressive nature of our framework, this enables token-level optimization of generation behaviors, aligning output quality directly with multimodal guidance.

We apply the vanilla GRPO implementation from the `trl` (von Werra et al.) library. The model is fine-tuned for 600 steps with a batch size of 12 and a group size of 8, resulting in 96 sampled generations per update. The training corpus includes 73K multimodal samples from UNO-1M (Wu et al., 2025b) and 60K text-to-image samples from BLIP-3O (Chen et al., 2025a). KL regularization is applied toward a frozen reference policy to maintain training stability.

We employ a composite reward that integrates multiple complementary signals reflecting perceptual quality, aesthetic appeal, and semantic fidelity. Specifically, for text-to-image (T2I) generation, the reward combines *HPSv2*, *Aesthetic*, *PickScore*, and *QwenVL-7B* scores. For multimodal generation, we additionally include a *CLIP-image* consistency score computed against the reference image. We utilize QwenVL-7B as a unified vision-language reward model to assess generation quality across T2I and multimodal settings.

Our GRPO-based fine-tuning effectively bridges autoregressive modeling and multimodal RL. By integrating structured, interpretable reward signals, we enable direct policy optimization toward higher visual quality and semantic alignment.

(1) Text-to-Image Evaluation Prompt

Instruction: You are a strict **reward model** for text-to-image (T2I) evaluation. Your goal is to produce an objective visual description and four quantitative sub-scores, with a total score in [0,4].
Inputs:

- Text prompt: `"{PROMPT}"`
- Generated image: `<image>`

Evaluation Procedure:

1. **Describe the generated image only** — objects, colors, lighting, composition, perspective, and visible defects. Do *not* infer from the prompt.
2. Provide four sub-scores (each $\in$ [0,1.0]): (A) Prompt/Category Alignment, (B) Completeness & Composition, (C) Realism & Visual Quality, (D) Artifact/Defect Freedom.
3. Compute the final score: raw $= A + B + C + D$; if any sub-score ≤ 0.3, cap final = 1.0. Otherwise, final = raw (rounded to two decimals).

Output: JSON with fields `{"description": "...", "scores": {"A":..,"B":..,"C":..,"D":..}, "final": ...}`

(2) Multimodal Generation Evaluation Prompt

Instruction: You are a strict **reward model** for subject-driven image generation with a reference image. Use the full [0,5] range and emphasize identity preservation and artifact freedom.
Inputs:

- Reference image: `<image>`
- Subject name: `"{SUBJECT}"`
- Text prompt: `"{PROMPT}"`
- Generated image: `<image>`

Evaluation Procedure:

1. Describe the generated image only, without inference.
2. Provide three sub-scores: (A) Instruction Following (0–0.6), (B) Identity Preservation vs. Reference (0–3.2), (C) Artifact/Defect Freedom (0–1.2).
3. Compute the final score: raw $= A + B + C$; if A $\leq$ 0.1, B $\leq$ 0.3, or C $\leq$ 0.3, cap final = 1.0. Otherwise, final = raw (rounded to two decimals).

Output: JSON with fields `{"description": "...", "scores": {"A":..,"B":..,"C":..}, "final": ...}`

C Text-to-Image Generation Evaluation

We evaluate the performance of our model on text-to-image (T2I) generation using the GenEval (Ghosh et al., 2024) benchmarks. Results are reported in Table 6.

Since MENTOR is built upon LLaMaGen—a relatively weaker autoregressive generator—its standalone T2I performance is inferior to earlier diffusion-based models such as LDM and SDv1.5. This is expected, as models based on more advanced generators (e.g., SDXL, SD3) such as KOSMOS-G and Dream Engine consistently outperform ours in conventional T2I metrics.

Nevertheless, MENTOR demonstrates strong performance in multimodal image generation tasks. Thanks to our proposed autoregressive architecture and a two-stage multimodal-conditioned tuning strategy, MENTOR effectively integrates both visual and textual modalities during generation. This synergistic design compensates for its weaker generation core, enabling MENTOR to surpass more powerful T2I models in multimodal settings, as shown in Table 3. We anticipate that incorporating stronger base generators will further improve performance. Despite its current limitations, our results suggest that MENTOR presents a promising and efficient alternative to diffusion-based methods in multimodal scenarios.

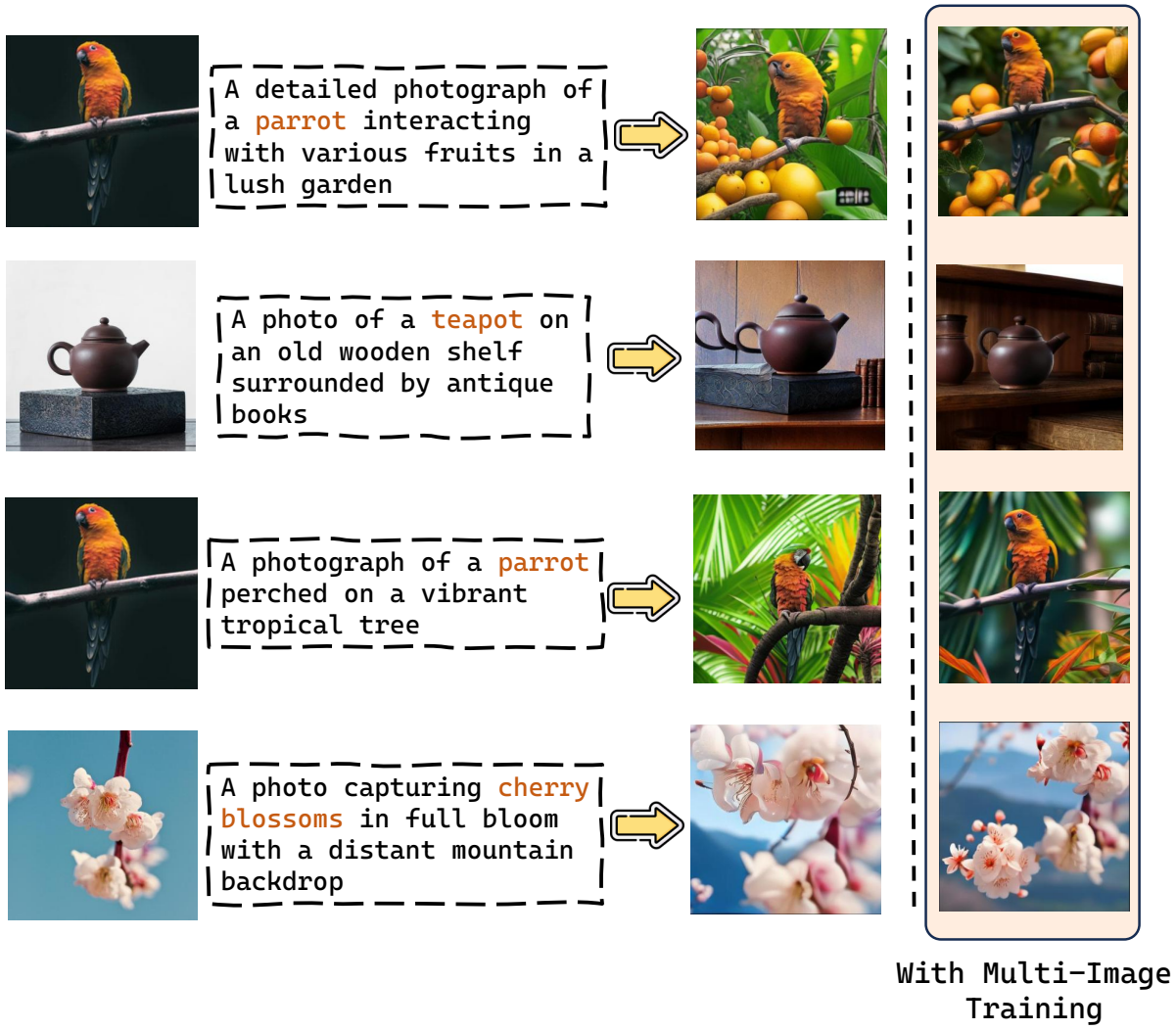

Figure 7: Qualitative assessment demonstrating improved preservation of visual details by MENTOR following multi-image training.

D Data Construction and Formation

Data Formation Table 7 summarizes the datasets utilized in our two-stage training framework. Each stage is designed to progressively enhance distinct capabilities of the model using a diverse collection of multimodal data sources. In total, approximately 3 million samples are employed, with Stage 1 comprising around 2.5 million samples and Stage 2 involving 1.3 million samples, including an overlap of roughly 800k examples.

The dataset is constructed from a combination of open-source resources, such as Midjourney (Emporium, 2024) and CC12M (Changpinyo et al., 2021), along with synthetic data generated via publicly available text-to-image (T2I) models, including Flux.1 (BlackForestLabs, 2024) and Stable Diffusion v3.5 (Esser et al., 2024).

Stage 1 focuses on establishing foundational multimodal alignment capabilities. Specifically, it includes 700k T2I samples from Midjourney (Emporium, 2024), 180k image reconstruction samples also from Midjourney, and 1.6M object segmentation samples generated through our pipeline.

Stage 2 fine-tunes the model with 1.3 million samples. This includes 600k T2I samples—200k from Midjourney and 400k synthesized using open-source T2I models such as Flux.1 (BlackForestLabs, 2024) and Stable Diffusion v3.5 (Esser et al., 2024). Additionally, we include 150k object segmentation samples and 150k image recovery samples, all derived from synthetic data using segmentation masks. Background images for the image recovery task are randomly selected from CC12M (Changpinyo et al., 2021).

We further incorporate 400k subject-driven image generation samples from Subject200k (Tan

Table 6: GenEval benchmark results for text-to-image generation, classifying methods as either autoregressive or diffusion-based models. Due to our method's model size and suboptimal generators, we experience poor performance in text-to-image generation.

	Method	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall
Autoregressive	Chameleon (Team, 2024b)	-	-	-	-	-	-	0.39
	LWM (Liu et al., 2024b)	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	LlamaGen (Sun et al., 2024b)	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	Show-o (Xie et al., 2024)	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	Emu3-Gen (Wang et al., 2024c)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	Janus (Wu et al., 2024b)	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	MENTOR	0.87	0.47	0.16	0.65	0.11	0.17	0.40
Diffusion	LDM (Rombach et al., 2022a)	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 (Rombach et al., 2022a)	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-α (Chen et al., 2023a)	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 (Rombach et al., 2022a)	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 (Ramesh et al., 2022)	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	SDXL (Podell et al., 2024)	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 (Betker et al., 2023)	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SDv3 Medium (Esser et al., 2024)	0.98	0.74	0.63	0.67	0.34	0.36	0.62
	Flux.1 Dev (BlackForestLabs, 2024)	0.98	0.81	0.74	0.79	0.22	0.45	0.66
	Dream Engine (Chen et al., 2025b)	1.00	0.94	0.64	0.81	0.27	0.49	0.69

Table 7: Details on dataset used in the two-stage training.

Stage	Data Source	Task	Number of Samples
1	Midjourney(Emporium, 2024)	Text to Image Generation	700k
	Midjourney(Emporium, 2024)	Image Reconstruction	180k
	Synthetic Data	Object Segmentation	1.6M
2	Midjourney(Emporium, 2024), Synthetic Data	Text to Image Generation	600k
	Synthetic Data	Object Segmentation	150k
	Synthetic Data, CC12M(Changpinyo et al., 2021)	Image Recovery	150k
	Subject200k(Tan et al., 2025)	Subject-driven Generation	400k

et al., 2025). These samples are re-captioned using Qwen2-VL (Wang et al., 2024a) to extract subject-relevant text and generate comprehensive image descriptions. To enrich the training set, we reverse the input-output image pairs, effectively doubling the usable data to 400k samples.

Data Construction To support the large-scale training required for our two-stage paradigm, we developed an automated pipeline for generating high-quality multimodal training data, as shown in Figure 8. This pipeline combines open-source image datasets with state-of-the-art vision-language models (VLMs) and segmentation models, enabling the construction of richly annotated image-text pairs with multiple segmented foreground objects without manual labeling:

- **Captioning and Object Extraction:** A VLM is queried to generate a comprehensive caption describing prominent elements in the image, followed by extracting a list of concrete, distinct, and segmentable objects. This ensures that the generated data focus on tangible visual entities.
- **Spatial Grounding:** For each extracted object, the VLM is queried again to identify its spatial location within the image, returning both a tight bounding box and several representative 2D key-points. These spatial cues constrain the region of interest for subsequent segmentation, improving accuracy and reducing background noise.
- **Segmentation:** A segmentation model is employed to extract object masks from the image, guided by the generated bounding boxes and key-points. This step produces high-quality masks that are both semantically aligned with the object labels and spatially accurate.

By applying this automated process to a large corpus of open-source images, we construct a di-

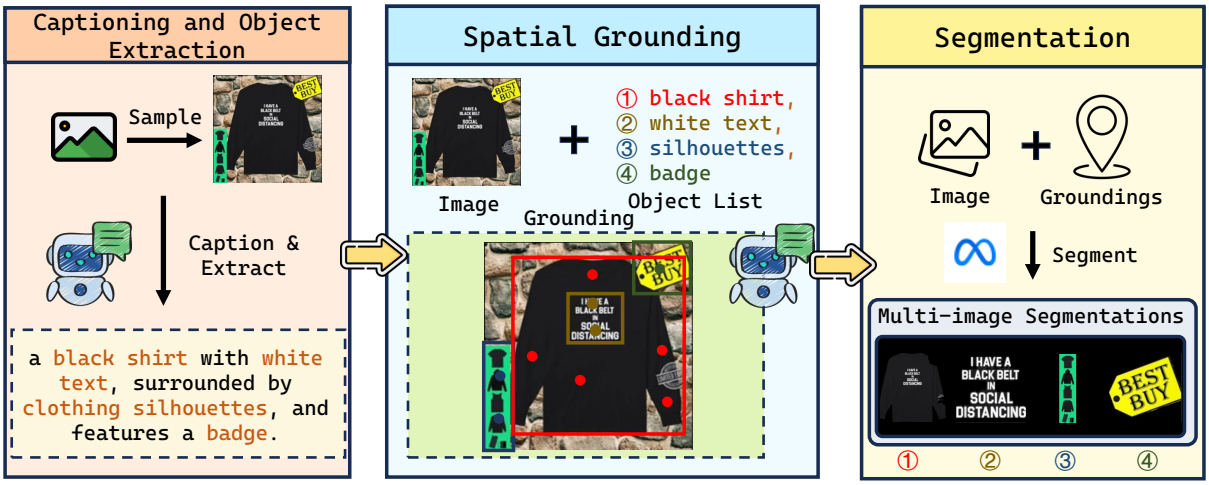

Figure 8: Illustration of the automatic data generation pipeline.

verse multimodal dataset comprising captioned images annotated with multiple precisely segmented objects. This dataset forms a critical component of our training setup, particularly enabling the object segmentation and image recovery tasks in our training paradigm.

E Experiment Details

E.1 Benchmark Details

DreamBench++ Data Organization. DreamBench++ (Peng et al., 2025) comprises 150 high-quality reference images, sourced from Unsplash, Rawpixel, and Google Images, encompassing a balanced mix of subjects. These are evenly divided into three broad categories: *objects*, *living subjects* (humans and animals), and *styles* (illustrative, painterly, etc.), ensuring visual and conceptual diversity.

In total, DreamBench++ offers **1,350 prompts** (150 × 9), representing a substantial scale-up over the original DreamBench (30 subjects × 25 prompts). Relative to DreamBench, the dataset is **5× larger in subjects** and **54× larger in prompts**, enabling broader evaluation of generative performance.

Evaluation Metric. DreamBench++ adopts an automatic, GPT-4o-based evaluation protocol designed to closely mirror human judgment. Each generated image is assessed against both its reference image and its corresponding prompt, using two complementary axes:

- **Concept Preservation (CP):** Measures fidelity between the generated image and the reference. Key attributes include shape, color, texture, and facial details.
- **Prompt Following (PF):** Evaluates how well the generation aligns with the prompt in terms of relevance, accuracy, completeness, and contextual appropriateness.

Each axis is scored on a **five-level ordinal scale** from 0 (Very Poor) to 4 (Excellent), avoiding the complexity and bias of pairwise comparisons.

Metric Justification. Following DreamBench++, we adopt **CP·PF** as the primary score for balanced performance and **CP/PF** as an indicator of modality overfitting. *Why CP·PF instead of CP+PF?* The additive form can assign high scores to models excelling on only one modality (e.g., CP≈1, PF≈0.1 still yields a large sum). In contrast, CP·PF approaches zero when *either* metric is poor, thus more sensitively reflecting imbalanced multimodal control. *Why CP/PF as an overfitting indicator?* On DreamBench++, the dominant failure mode is *image-dominant* behavior—copying the reference while ignoring text. CP/PF>1 directly captures this (e.g., Lumina-mGPT and Unified-IO2 have CP/PF>3). Using PF/CP instead would flag text-dominant failures, which are less common in this benchmark.

DreamBench The original DreamBench (Ruiz et al., 2023) dataset consists of 30 subjects, each paired with 25 prompts, totaling 750 prompt-image pairs. It serves as a foundational benchmark for evaluating personalized image generation models, focusing on the model's ability to maintain subject identity across diverse prompts.

E.2 Baselines

We compare our method against various baselines, categorized as follows:

Figure 9: Qualitative comparison of image reconstruction results using MENTOR.

- **Textual Inversion** (Gal et al., 2022) learns a new word embedding to represent a specific concept, enabling personalized image generation by incorporating the new token into prompts. It requires a few images of the subject and fine-tunes the embedding without altering the base model weights.
- **DreamBooth** (Ruiz et al., 2023): DreamBooth fine-tunes a pre-trained text-to-image model to bind a unique identifier with the subject's visual concept, allowing for personalized generation. It requires several images of subject and modifies model weights to capture subject-specific details.
- **BLIP-Diffusion** (Li et al., 2023a): This approach introduces a pre-trained multimodal encoder to provide subject representations for the diffusion generator, enabling controllable multimodal image generation.
- **KOSMOS-G** (Pan et al., 2024): KOSMOS-G is a multimodal large language model designed for zero-shot image generation from interleaved vision-language inputs, including multiple images and text. It aligns the output space of a transformer-based causal language model with a diffusion-based image decoder using a lightweight AlignerNet and compositional instruction tuning. This architecture enables KOSMOS-G to perceive complex multimodal prompts and generate coherent, subject-driven images without modifying the base image decoder.
- **Emu2** (Sun et al., 2024c): Emu2 is a 37-billion-parameter generative multimodal model trained on large-scale multimodal sequences with a unified autoregressive objective. It exhibits strong in-context learning abilities for various multimodal tasks, including visual prompting and object-grounded generation.
- **IP-Adapter** (Ye et al., 2023): IP-Adapter is a lightweight adapter that enables image prompt capability for pre-trained text-to-image diffusion models. It integrates image features into the generation process without modifying the base model, supporting flexible and efficient image-to-image generation.
- **DreamEngine** (Chen et al., 2025b): DreamEngine is a unified framework that integrates multimodal encoders with diffusion models through a two-stage training approach, enabling advanced text-image interleaved control and achieving state-of-the-art performance in generating images with complex, concept-merged inputs.
- **Unified-IO 2** (Lu et al., 2023): Unified-IO 2 is an autoregressive multimodal model capable of understanding and generating images, text, audio, and actions. It tokenizes various modalities into a shared semantic space and processes them with a single encoder-decoder transformer. Trained from scratch on a large multimodal pre-training corpus and fine-tuned on an ensemble of 120 datasets, Unified-IO 2 achieves state-of-

Figure 10: Image reconstruction results of MENTOR *without* alignment tuning.

the-art performance on the GRIT benchmark and strong results across more than 35 benchmarks.

- **Lumina-mGPT** (Zhuo et al., 2024): Lumina-mGPT is a multimodal autoregressive model designed for flexible photorealistic text-to-image generation. It employs a pretrained decoder-only transformer as a unified framework for modeling multimodal token sequences. Through multimodal Generative PreTraining (mGPT) and subsequent Flexible Progressive Supervised Finetuning (FP-SFT) and Omnipotent Supervised Finetuning (Omni-SFT), Lumina-mGPT demonstrates versatile multimodal capabilities, including visual generation tasks, controllable generation tasks and vision-language tasks.

F Qualitative Study

F.1 Image Reconstruction

As illustrated in Figures 9 and 10, MENTOR demonstrates strong image reconstruction capabilities following two-stage training. Notably, it is able to effectively reconstruct input images and preserve fine-grained visual details, even when input images are of low resolution (224×224) and outputs are generated at 512×512 resolution.

In contrast, when alignment tuning is omitted, although the model benefits from the pretrained multimodal encoder and the proposed architecture, it tends to treat the input image as a visual prompt akin to a caption. As shown in Figure 10, this leads to outputs that resemble descriptive interpretations of the input rather than faithful reconstructions. Consequently, visual fidelity and spatial consistency degrade significantly without alignment tuning.

F.2 Text-guided Image Segmentation

We evaluate MENTOR on the DreamBench++ benchmark to assess its performance in text-guided image segmentation. As demonstrated in Figure 11, MENTOR successfully identifies and segments visual concepts corresponding to the given textual prompts. These results highlight the model's ability to generalize across tasks and showcase its robust multimodal understanding and generation.

Open-Vocabulary Segmentation. To further assess segmentation capability, we evaluate MENTOR on ADE20K (A-150), a standard open-vocabulary segmentation benchmark. Notably, we perform zero-shot evaluation directly after Stage-1 pretraining, without any task-specific fine-tuning on these datasets. Due to the fixed output resolution of our generator, all input images are resized to 512×512 for inference.

As shown in Table 8 and Figure 12, MENTOR achieves competitive performance on the benchmark compared with the traditional open-vocabulary segmentation methods in a zero-shot manner. Compared with these methods that require massive task-specific pretraining, MENTOR shows competitive performance after Stage-1 training in a zero-shot manner. These results demonstrate that our two-stage training effectively learns

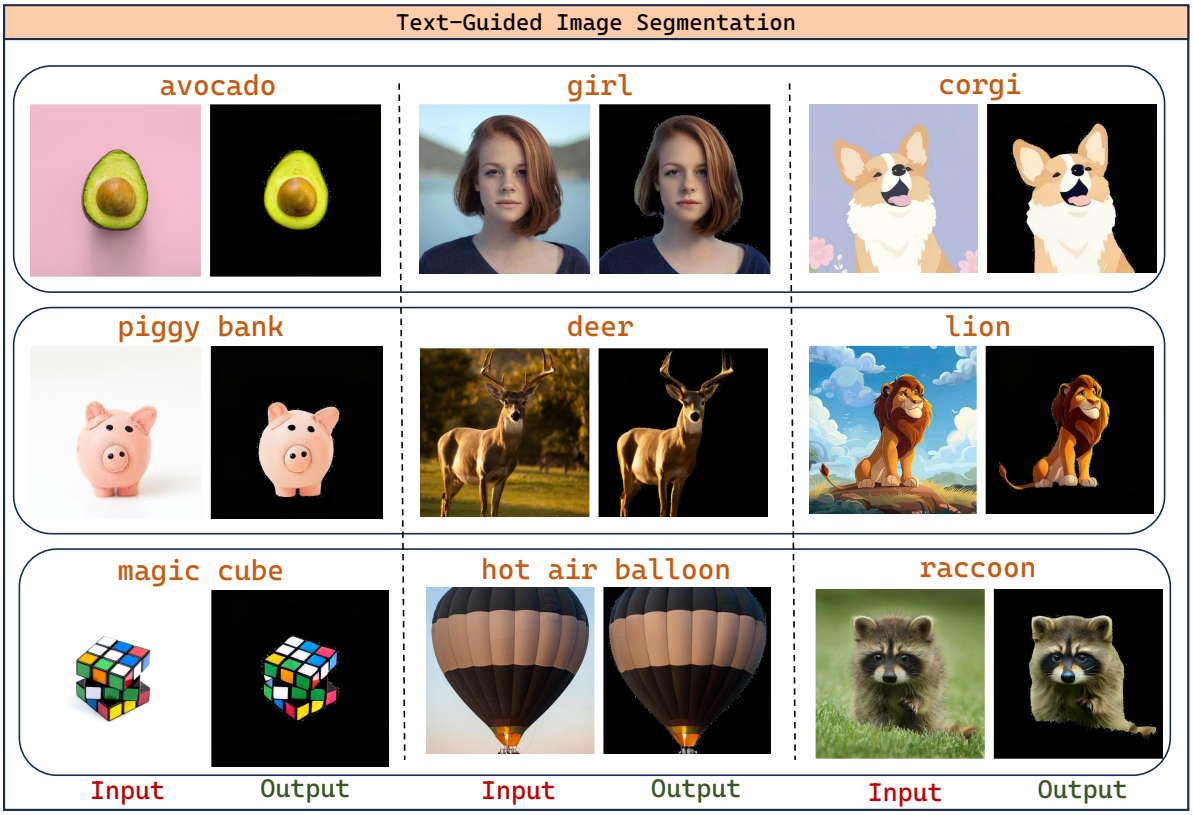

Figure 11: Qualitative results of text-guided image segmentation using MENTOR.

Table 8: Zero-shot open-vocabulary segmentation on ADE20K-150 (Zhou et al., 2017). We report the mIoU score as the metric.

Dataset	LSeg+ (Li et al., 2022)	OpenSeg (Li et al., 2022)	SimBaseline (Li et al., 2022)	ZegFormer (Li et al., 2022)	SimSeg (Li et al., 2022)	DreamLIP (Li et al., 2022)	MENTOR
ADE20K-150	13.00	15.30	15.30	16.40	20.50	17.10	19.91

to balance visual and textual guidance, enabling the model to generate accurate segmentation masks for novel categories. While performance is constrained by the current backbone capacity and data scale, these findings highlight the extensibility and promising potential of our framework for broader vision-language tasks.

F.3 Multi-Image Generation

We evaluate MENTOR on multi-image generation tasks using the X2I dataset (Xiao et al., 2024b). As shown in Figure 14, the model is capable of generating visually consistent outputs conditioned on the multi-image inputs. The generated images reflect coherent semantics, style, and layout across the samples.

F.4 Multimodal In-Context Image Generation

To assess MENTOR's few-shot generalization capabilities, we evaluate it on the multimodal in-context image generation task using the X2I-ICL dataset (Xiao et al., 2024b). As illustrated in Figure 15, MENTOR learns to synthesize images that follow the stylistic patterns demonstrated in the in-context examples. This indicates its capability to infer complex visual trends and align generation with image context.

G Versatility Across Different Multimodal Tasks

To assess the broad applicability of our proposed framework, we evaluate MENTOR across a diverse set of multimodal generation tasks, including text-guided image segmentation, subject-driven image generation, multi-image generation, and multimodal in-context learning. For each task, we apply supervised fine-tuning where necessary, ensuring robust generalization while maintaining architectural consistency.

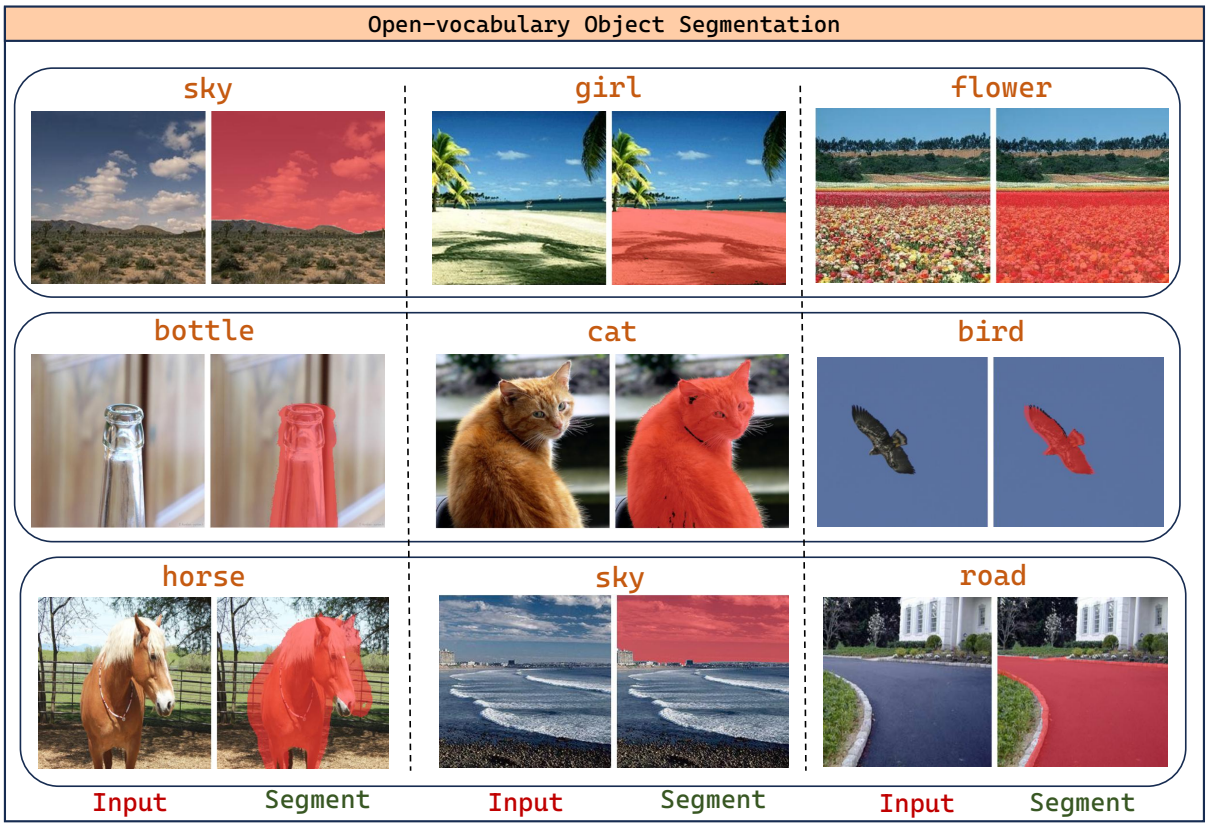

Figure 12: Qualitative results of open-vocabulary segmentation using MENTOR on ADE20K (Zhou et al., 2017), PASCAL Context (Mottaghi et al., 2014), and PASCAL VOC (Everingham and Winn, 2011) benchmarks. The generated segmentation masks are visualized in red for clarity.

Image Segmentation. We evaluate this task directly after Stage 1 training, without additional fine-tuning. The model demonstrates strong object localization and mask precision from prompt-aligned inputs, confirming the effectiveness of the proposed training pipeline and segmentation-aware data construction process.

Subject-driven Image Generation. This task is evaluated using the model at the end of Stage 2. No additional task-specific tuning is applied. The model successfully generates high-fidelity, identity-preserving images consistent with subject descriptors.

Multi-Image Generation. We fine-tune the Stage 2 model on a subset of X2I-subject-driven (Xiao et al., 2024b) dataset for two additional epochs using a reduced learning rate of 5×10^{-5}. All other optimization settings remain consistent with Stage 2. The dataset is split into disjoint training and test sets, and quantitative results are reported on the test split. The model learns to generate visually diverse yet semantically aligned images for the same input.

Multimodal In-Context Learning. We fine-tune the model for 10 epochs on the X2I-ICL dataset (Xiao et al., 2024b), which features sequences of input-output pairs for in-context generalization. We use a learning rate of 5×10^{-5} and ensure a strict train-test separation. The model adapts to context examples and generates new samples following the observed patterns, showing strong in-context learning performance without explicit prompt engineering.

Conclusion. The qualitative results presented in Section F confirm the versatility of MENTOR across a wide range of tasks. Notably, the model adapts to each task without architectural modifications, requiring only lightweight fine-tuning.

H Human Evaluation

To validate the reliability of GPT-based evaluation metrics used in DreamBench++, we conduct a human study on 50 randomly sampled images from our model's outputs. Three independent annotators rate each image on Concept Preservation (CP) and Prompt Following (PF) using the same five-level

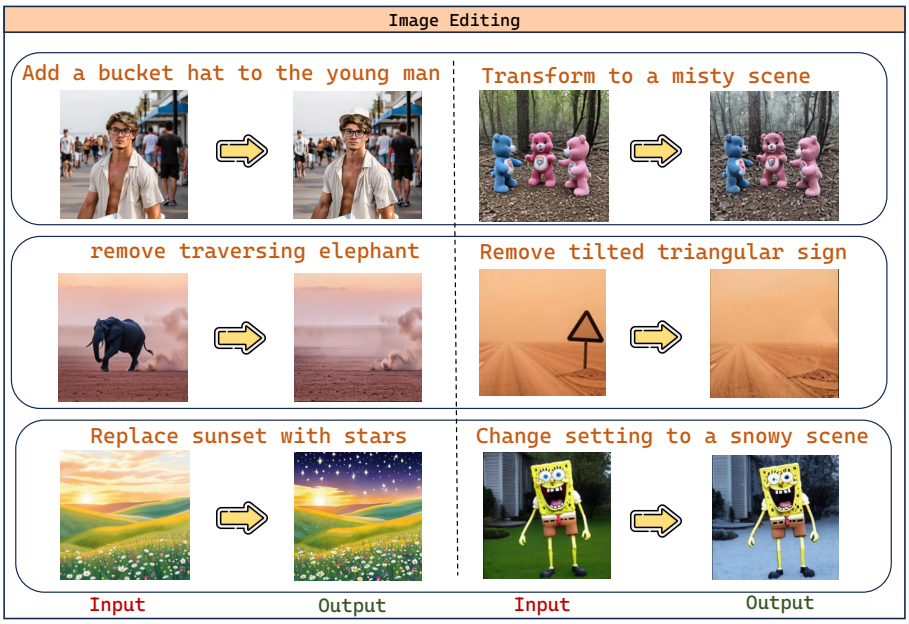

Figure 13: Qualitative examples from Image Editing. The model is finetuned on 1M samples from ImgEdit (Ye et al., 2025) and OmniEdit (Wei et al., 2024) dataset. Although restricted by the performance of image generator, MENTOR can generate visually consistent edits.

ordinal scale (0–4) as the automatic evaluation.

We report two complementary measures of reliability:

- **Inter-annotator agreement:** Measured by Krippendorff's α, we obtain $\alpha = 0.76$ for CP (substantial agreement) and $\alpha = 0.62$ for PF (moderate agreement). These values indicate consistent human judgment across annotators.
- **Human-metric consistency:** Defined as the proportion of cases where the GPT-based score matches the majority human judgment, this metric reaches 84% for CP and 94% for PF. These figures align closely with the consistency rates reported in DreamBench++ (83.31% for CP, 98.17% for PF).

These results confirm that the GPT-based metrics serve as reliable proxies for human evaluation, supporting their use as the primary evaluation protocol in our experiments.

I CP–PF Trade-off Analysis

A central challenge in multimodal-conditioned image generation is balancing fidelity to the reference image (Concept Preservation, CP) against adherence to the text prompt (Prompt Following, PF). In this section, we conduct systematic experiments to analyze the CP–PF trade-off by varying two key factors: (1) Stage-2 task mixing ratios during training, and (2) classifier-free guidance (CFG) scale at inference.

I.1 Effect of Task Weight Distribution

We investigate how the composition of training tasks in Stage-2 affects the balance between concept preservation and prompt following. Stage-2 comprises four task types with configurable sampling weights: text-to-image generation (T2I), object segmentation (Seg), image recovery (Rec), and subject-driven generation (Sub). We evaluate five representative weight configurations, as reported in Table 9.

The results reveal a systematic trade-off between concept preservation and prompt following:

- **Image-heavy:** Increasing weights for image-centric tasks (Seg, Rec, Sub) improves CP from 0.557 to 0.609 but reduces PF from 0.840 to 0.741.
- **Text-heavy:** Emphasizing T2I (weight = 0.80) yields higher PF (0.915) at the cost of significantly degraded CP (0.383).
- **w/o Rec:** Reducing image recovery weight primarily degrades PF and induces copy-paste behavior.
- **w/o Seg:** Minimizing segmentation weight mainly hurts CP, impairing fine-grained visual grounding.

These observations are consistent with our ablation

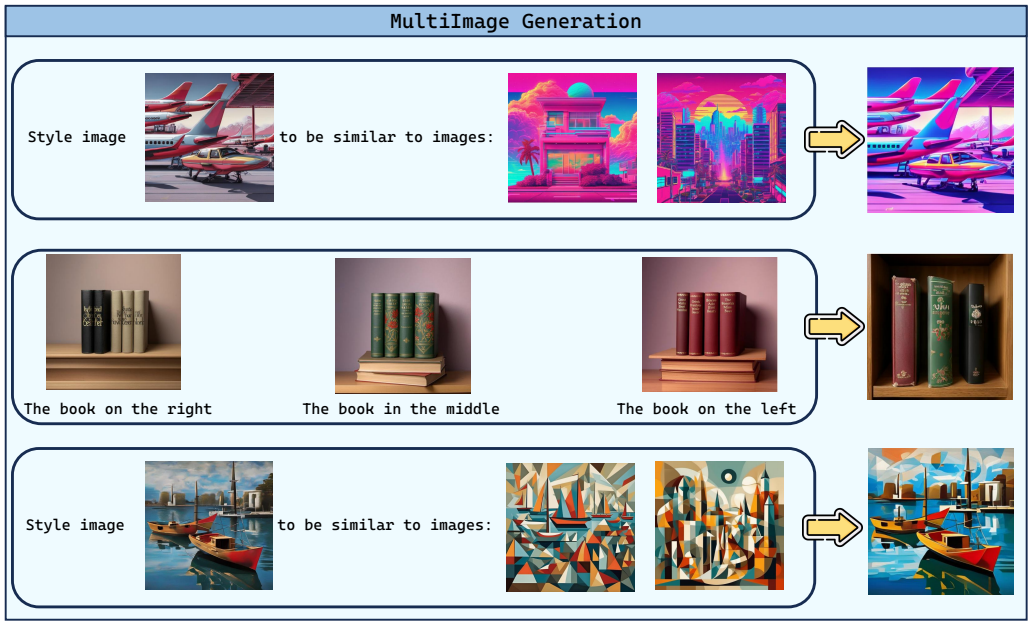

Figure 14: Qualitative results for multi-image generation.

Table 9: Effect of Stage-2 task weight distribution on DreamBench++. Left columns show sampling weights for each task type; right columns report evaluation metrics.

	Task Sampling Weights				Metrics		
Configuration	T2I	Seg	Rec	Sub	CP↑	PF↑	CP·PF↑
Default (Ours)	0.45	0.15	0.15	0.25	0.557	0.840	**0.468**
Image-heavy	0.20	0.25	0.25	0.30	0.609	0.741	0.451
Text-heavy	0.80	0.05	0.05	0.10	0.383	0.915	0.350
w/o Rec	0.45	0.25	0.05	0.25	0.588	0.666	0.392
w/o Seg	0.45	0.05	0.25	0.25	0.390	0.880	0.343

Table 10: Effect of CFG scale on DreamBench++ metrics. Higher CFG strengthens prompt adherence but may degrade concept preservation at extreme values.

CFG Scale	**3.5**	**5.5**	**7.5**	**9.5**
CP↑	0.532	0.551	0.557	0.481
PF↑	0.815	0.828	0.840	0.881
CP·PF↑	0.434	0.456	**0.468**	0.424

study (§3.3), confirming that the default configuration achieves the optimal balance as measured by CP·PF.

I.2 Effect of Classifier-Free Guidance Scale

We examine how classifier-free guidance (CFG) scale at inference time affects the CP–PF trade-off. Table 10 summarizes the results across four CFG values.

Increasing CFG leads to monotonic improvement in PF, as stronger guidance amplifies the influence of text conditioning. However, CP exhibits a non-monotonic pattern: it initially rises with CFG but drops sharply at CFG=9.5, likely due to over-amplification that distorts fine-grained visual features. The combined metric CP·PF peaks at CFG=7.5, which we adopt as the default following the DreamBench++ protocol.

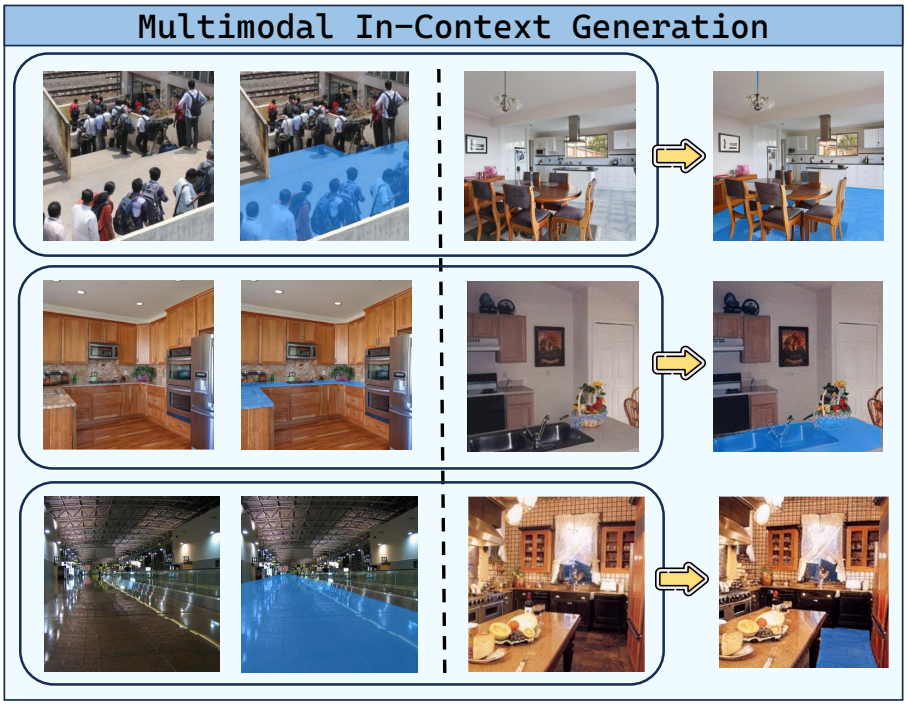

Figure 15: Qualitative examples from multimodal in-context image generation. The model adapts to patterns in the visual context.