

# MENTOR: Efficient Multimodal-Conditioned Tuning for Autoregressive Vision Generation Models

Haozhe Zhao<sup>1\*</sup>, Zefan Cai<sup>2\*</sup>, Shuzheng Si<sup>3</sup>, Liang Chen<sup>4</sup>,  
Jiuxiang Gu<sup>5</sup>, Wen Xiao<sup>6</sup>, Junjie Hu<sup>2</sup>

<sup>1</sup>University of Illinois Urbana-Champaign <sup>2</sup>University of Wisconsin-Madison

<sup>3</sup>Tsinghua University <sup>4</sup>Peking University <sup>5</sup>Adobe Research <sup>6</sup>Microsoft

haozhez6@illinois.edu

mentor.github.io

## Abstract

Recent text-to-image models produce high-quality results but still struggle with precise visual control, balancing multimodal inputs, and demanding extensive training for complex multimodal image generation. To address these limitations, we propose **MENTOR**, a novel autoregressive (AR) framework for efficient **M**ultimodal-**co**ndition**E**d **tu**Ning for **au**Toreg**R**essive multimodal image generation. MENTOR combines an AR image generator with a two-stage training paradigm, enabling fine-grained, token-level alignment between multimodal inputs and image outputs—without relying on auxiliary adapters or cross-attention modules. Central to our method is the two-stage training paradigm: (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 balance model’s integration of multimodal inputs and enhance generation controllability. Extensive experiments demonstrate that, despite a modest model size, suboptimal base components, and limited training resources, MENTOR achieves a strong balance between concept preservation and prompt following on DreamBench++ benchmark, outperforming competitive baselines. Additionally, our method also delivers superior image reconstruction fidelity, broad adaptability across multimodal tasks, and an efficient training budget compared to diffusion-based counterparts. The dataset, code, and models are available in [github.com/HaozheZhao/MENTOR](https://github.com/HaozheZhao/MENTOR).

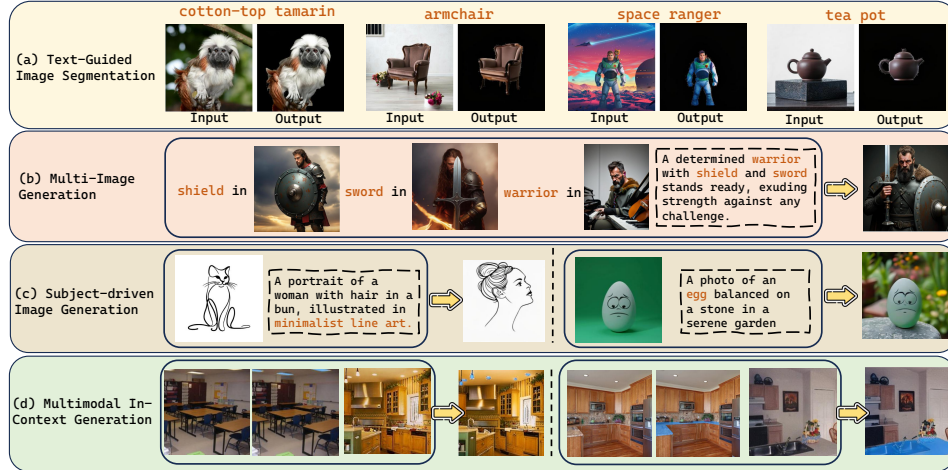


Figure 1: Versatile applications built on MENTOR after simply fine-tuning on corresponding datasets.

\*Equal contribution.

# 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 high-quality image generation, e.g., fine-grained control over generated images, the models need to seamlessly integrate multi-modal inputs, such as reference images with detailed instructions. This poses significant challenges for existing diffusion models that are predominantly focused on T2I generation. To address this, researchers have recently employed Large Multimodal Models (LMMs) to encode diverse inputs into unified embeddings compatible with diffusion models [Pan et al., 2024, Sun et al., 2024c]. While this approach aids in handling complex prompts for generation like interleaved image-text instructions [Sun et al., 2024c] or multimodal in-context learning [Pan et al., 2024, Ge et al., 2024], several key limitations persist when scaling to complex multimodal control:

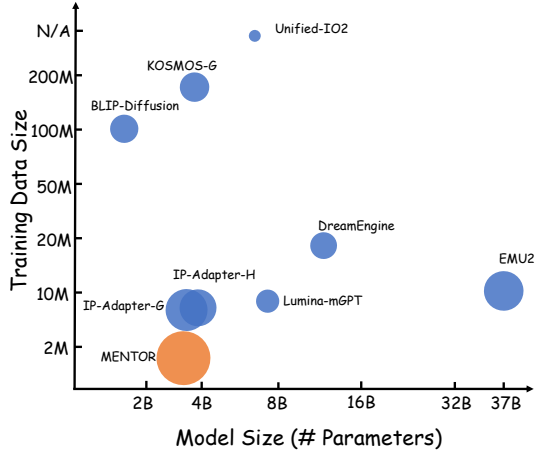


Figure 2: CP-PF score (circle size) of **MENTOR** and **other baselines** on DreamBench++. Model in lower left achieves the best efficiency.

**First**, the randomness of diffusion processes hinders precise, deterministic control, which is essential for high-fidelity tasks, like image reconstruction[Wang et al., 2025]. **Second**, balancing guidance from different modalities remains challenging. Existing methods frequently exhibit modality imbalance, often over-emphasizing one modality while neglecting others[Han et al., 2024]. For instance, IP-Adapter [Ye et al., 2023] and Lumina-mGPT [Zhuo et al., 2024], conditioned on text and image features, may overly favor image inputs. This imbalance stems 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 et al., 2024]. **Third**, many diffusion-based methods [Pan et al., 2024, Sun et al., 2024c], particularly those with auxiliary alignment components (e.g., learned adapters [Pan et al., 2024], regression heads [Sun et al., 2023, 2024c], specialized embeddings [Ge et al., 2023]), demand large-scale training [Sun et al., 2024c, Pan et al., 2024, Li et al., 2023a, Ge et al., 2024], incurring substantial computational costs. These challenges raise a critical question: *Is there a more efficient and controllable paradigm for balancing complex multimodal conditions in image generation?*

To address aforementioned limitations, we propose **MENTOR**, a straightforward and efficient autoregressive (AR) framework for controllable multimodal image generation. Unlike diffusion models that rely on complex cross-attention layers for multimodal conditioning and demand extensive training resources [Ge et al., 2024, Sun et al., 2024c, Pan et al., 2024, Li et al., 2023a], MENTOR leverages a unified transformer architecture that directly aligns multimodal inputs with output image tokens. This design inherently simplifies architecture and training, reduces the need for auxiliary components for alignment, and significantly lowers training requirements. Our framework employs a multimodal encoder to project multimodal inputs into a unified representation. An AR transformer decoder then generates image tokens deterministically, conditioned on this representation. To ensure effective and balanced modality integration, which is curial for multimodal image generation[Han et al., 2024], we adopt a two-stage training paradigm: (1) pretraining on image reconstruction and object segmentation to enable robust pixel-level and semantic alignment, followed by (2) multimodal instruction tuning with diverse tasks, such as image recovery and subject-driven generation, explicitly training the model to effectively leverage and balance varied multimodal inputs for coherent generation.

Notably, despite its simplicity and the use of suboptimal checkpoints, MENTOR achieves state-of-the-art(SOTA) performance on the challenging DreamBench++ benchmark—using 10× less training data than leading baselines. It outperforms resource-intensive baselines with powerful generators like SDXL [Podell et al., 2023] and SD3 [Esser et al., 2024] by 30% while dramatically reducing computational and data demands. Controlled experiments demonstrate the advantages of MENTOR over diffusion-based methods in multimodal fidelity, training efficiency, and generation controllability.

Overall, our contributions are as follows: (1) A novel autoregressive framework for efficient multimodal generation, achieving, achieving SOTA performance; (2) A two-stage training strategy for multimodal-conditioned tuning, enabling robust alignment and balanced modality integration with significantly reduced training cost; (3) Extensive experiments demonstrating the superior efficiency, controllability, and fidelity of MENTOR as a compelling alternative to diffusion-based methods.

## 2 Method

This section details an efficient autoregressive framework for controllable multimodal image generation, achieving precise image control and balancing guidance across multiple modalities with minimal cost. § 2.1 describes the autoregressive training objective for our framework. Next, § 2.2 presents a straightforward yet efficient model architecture, detailing how the framework accommodates multimodal inputs and supports autoregressive generation. Crucially, § 2.3 introduces a two-stage training paradigm aimed at balancing the influences of different modalities during image generation. Finally, § 2.4 describes our automated pipeline for scalable multimodal data curation.

### 2.1 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  $\phi$  first encodes  $\mathbf{c}^{(0)}$  and subsequently uses an MLP layer to project them into space of the image decoder to form a unified representation  $\mathbf{h}$ :

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

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

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

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

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

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

### 2.2 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. A lightweight projection layer bridges the encoder’s output to the decoder’s input embedding space, enabling seamless integration between the two components.

**Multimodal Encoder** The multimodal encoder integrates multimodal inputs from frozen pretrained vision ( $\phi_V$ ) and language ( $\phi_L$ ) encoders into a shared latent space. This module projects visual features from  $\phi_V$  into  $\phi_L$ ’s embedding space using a lightweight connector module ( $\psi$ ), yielding a unified multimodal representation  $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_M)$ , where  $\mathbf{h}_j \in \mathbb{R}^d$ . Meanwhile, a critical challenge arises from the visual tokenization, where a vision encoder produces hundreds of tokens per image. It inflates the context length and leads to substantial computational costs, especially in multi-image scenarios. Compressing visual tokens can mitigate costs, but may risk losing fine-grained visual information. To navigate this trade-off, two architectures for  $\psi$  are investigated:

- **MLP-based Projection:** Inspired by Liu et al. [2023], we adopt a multi-layer perceptron that directly projects visual tokens to maintain detailed visual semantics with minimal information loss.

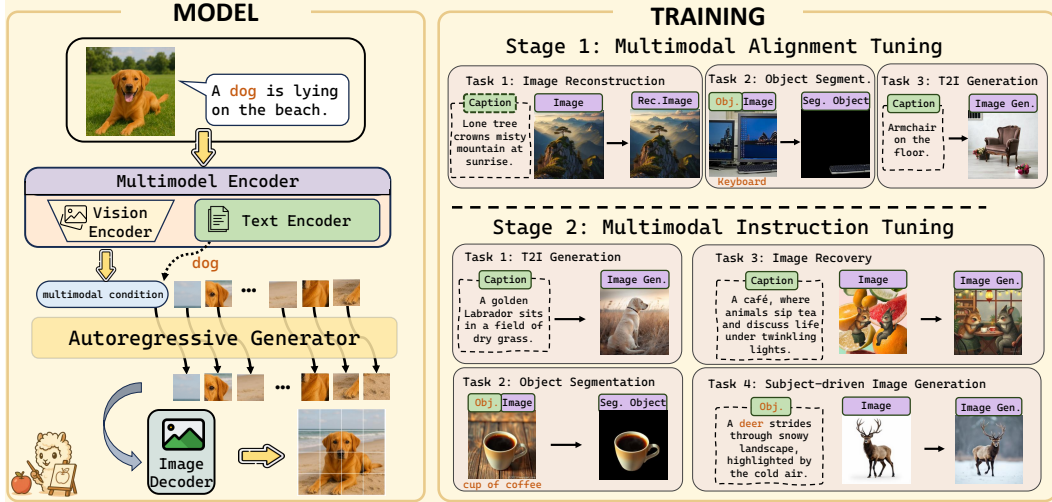


Figure 3: Overview of MENTOR. **Left** panel illustrates the model structure, where visual and textual inputs are encoded into a unified latent to guide autoregressive image generation. **Right** panel highlights the two-stage training paradigm: (1) **Multimodal Alignment Tuning**, enabling pixel and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

- **Query-based Token Distillation:** We use a Query-Former Li et al. [2023b] to compress lengthy visual token sequences into a fixed-size representation. To guide the distillation, we condition the Query-Former on textual queries [Li et al., 2023a, Dai et al., 2023] that highlight important concepts in images. It aims to guide the model to selectively extract semantically relevant visual features based on textual queries.

**Autoregressive Generation Decoder** A transformer-based autoregressive decoder generates a image token sequence  $\mathbf{y} = (y_1, \dots, y_L)$  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.

### 2.3 Two-Stage Training Paradigm

As highlighted in the introduction, effectively aligning disparate modalities and balancing their influence are crucial challenges for multimodal conditional generation. Our carefully designed 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** While the initial connector training provides preliminary semantic alignment, we observed that the model primarily interprets visual inputs semantically like text captions, neglecting crucial visual and spatial details necessary for precise image generation. This stage is dedicated to explicitly enhancing both pixel and semantic-level modality alignment and promoting the utilization of visual information. In Stage 1, we employ three complementary tasks:

- **Image reconstruction**, where models must faithfully reconstruct an input image conditioned on itself, with the corresponding caption randomly provided or omitted, reinforcing pixel-level fidelity.
- **Object segmentation**, where models are given an input image and a target object label and must generate an end-to-end segmented figure for that object. This task compels the model to explicitly capture fine-grained visual details and spatial structures associated with specific semantic concepts.
- **T2I generation**, using image-caption pairs to preserve and reinforce foundational generative capabilities learned during pretraining of the decoder.

Importantly, incorporating the segmentation task alongside image reconstruction mitigates the risk of the model degenerating into a trivial copy-paste behavior, such as simply replicating the input. The segmentation objective compels the model to produce semantically meaningful and spatially precise visual outputs. This complementary effect has been further analyzed and validated in § 3.3.

**Stage 2: Multimodal Instruction Tuning** Stage 2 aims to endow the model with robust instruction-following and cross-modal reasoning capabilities for nuanced and controllable multimodal generation,

building upon the alignment and visual fidelity established in Stage 1. The model is expected to jointly attend to and integrate diverse input modalities in a balanced and controllable manner. To achieve this, we employ a multimodal instruction tuning strategy based on a carefully curated mixture of training tasks. Specifically, we reuse the **T2I generation** and **object segmentation** tasks from Stage 1, maintaining identical data and formulations. These tasks respectively reinforce the model’s ability to adhere to and utilize textual and visual modalities, helping preserve foundational skills and stabilize the training process. In addition, we introduce two novel tasks specifically designed to enhance instruction adherence and foster a balanced integration of multimodal inputs, preventing the model from over-emphasizing a single modality while neglecting others:

- **Image recovery**, where we synthetically distort images by rotating, resizing, and compositing segmented objects onto random backgrounds, then pair synthetic images with their original captions to create inputs. The model is then required to reconstruct the original image from the distorted input and corresponding caption. It compels the model to extract and integrate essential visual details from noisy or incomplete visual inputs while leveraging textual cues to infer and restore missing components, promoting robust multimodal reasoning and error-correction performance.
- **Subject-driven image generation**, where the model is conditioned on reference image, subject label and textual instruction to generate images. It requires the model to actively preserve the subject’s visual identity from the reference image while strictly adhering to the textual instructions for image generation. This task serves as a comprehensive end-to-end objective, fully exercising model’s cross-modal fusion and instruction-following abilities.

Overall, this training strategy—combining continued refinement of core capabilities with targeted instruction-based tasks—ensures the model to learn to integrate visual and textual information in a harmonious and controllable way. It mitigates over-reliance on certain modality and enables precise, controllable multimodal conditional generation, which are further discussed in § 3.3.

## 2.4 Data Construction

To support our two-stage training paradigm, we construct a large-scale multimodal dataset comprising approximately **3 million samples** across all training tasks. It integrates open-source resources, synthetic data, and automated annotations to ensure scalability, diversity, and strong task alignment. For **image reconstruction** and **T2I generation**, we collect image-text pairs from datasets like CC12M [Changpinyo et al., 2021] and Midjourney-Niji [Emporium, 2024]. To broaden domain coverage (e.g., human subjects, artistic scenes), we generate additional samples using T2I models like Flux.1 [BlackForestLabs, 2024] and Stable Diffusion v3.5 [Esser et al., 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]. Given an image, LMMs are queried to produce a comprehensive caption and extract a list of concrete, segmentable objects. For each object, LMMs predict its spatial location, providing a bounding box and a set of 2D keypoints, which guide the segmentation model in producing high-quality, semantically consistent masks. 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 construction and formation are detailed in Appendix D.

## 3 Experiments

### 3.1 Experimental Setup

**Implementation Details** The multimodal encoder is initialized using CLIP-Large-Patch14 [Radford et al., 2021] and FlanT5-XL [Chung et al., 2022], with a 224×224 image receptive field. The generator is initialized from LlamaGen-XL [Sun et al., 2024b](775M). A two-layer MLP serves as the projector. We freeze the encoder, training only the projector and generator for one epoch in Stage 1, then fine-tune the full model (except the vision encoder) for two more epochs in Stage 2. Training on 8 A100 GPUs (80 GB each) takes about 1.5 days. More details are in Appendix B.

**Benchmark & Metric.** We evaluate MENTOR on **DreamBench** [Ruiz et al., 2023] and **DreamBench++** [Peng et al., 2025] benchmarks. **DreamBench** employs CLIP and DINO scores to measure

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	0.36	0.31	0.36	0.38	0.67	0.69	0.44	0.62	0.24	0.61
DreamBooth	SD v1.5	-	860M	0.64	0.20	0.49	0.48	0.49	0.79	0.78	0.50	0.72	0.36	0.68
DreamBooth-L	SDXL v1.0	-	2.60B	0.75	0.31	0.54	0.72	0.60	0.90	0.90	0.75	0.87	0.52	0.69
Test-Time Tuning-Free Methods														
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	3.79
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	3.64
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	0.51	0.30	0.50	0.27	1.10
Kosmos-G	SD v1.5	200M	3.00B	0.62	0.63	0.46	0.57	0.54	0.48	0.62	0.41	0.51	0.28	1.06
IP-A-Plus ViT-H	SDXL v1.0	10M	3.00B	0.90	0.85	0.76	0.91	0.83	0.50	0.38	0.28	0.41	0.34	2.02
Emu2	SDXL v1.0	16M	37.00B	0.67	0.55	0.45	0.45	0.53	0.73	0.72	0.56	0.69	0.36	0.77
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	0.55	0.86	0.85	0.80	0.84	<b>0.47</b>	<b>0.65</b>

the fidelity to images and prompts. **DreamBench++** offers a scaled and diverse evaluation dataset and introduces a human-aligned automatic evaluation protocol using GPT-4o, addressing limitations of DreamBench evaluation. GPT-4o evaluator scores generations 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. Details can be found in the Appendix E.1.

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

- **Fine-tuning-based methods:** Textual Inversion [Gal et al., 2022], and DreamBooth [Peng et al., 2025], which fine-tune models with auxiliary control mechanisms on a subset of benchmark data.
- **Test Time Tuning-Free Methods:** Models pretrained on large-scale data and evaluated in a zero-shot. Diffusion-based models like BLIP-Diffusion [Li et al., 2023a], Emu2 [Sun et al., 2024c], variants of IP-Adapter [Ye et al., 2023], and DreamEngine [Chen et al., 2025], as well as autoregressive models like Unified-IO 2 [Lu et al., 2023] and Lumina-mGPT [Zhuo et al., 2024].

### 3.2 Main results

Table 3 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. MENTOR rivals fine-tuned methods like DreamBooth-LoRA, while significantly outperforming test-time tuning-free baselines such as Emu2 and DreamEngine. For instance, MENTOR surpasses Emu2 in CP·PF score by approximately 30%. A key strength of MENTOR is its ability to harmoniously integrate multimodal inputs. Several strong baselines, including some AR methods like Lumina-mGPT and Unified-IO2, exhibit very high Concept Preservation (CP) but suffer from extremely low Prompt Following (PF), resulting in high **CP/PF** scores. This indicates a tendency to over-rely on the reference image while neglecting textual instructions. In contrast, MENTOR delivers lowest CP/PF scores compared to all test-time tuning-free 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 SOTA performance with balanced performance, highlighting its efficiency and effectiveness. Furthermore, MENTOR remains highly competitive in size compared to larger counterparts like Emu2 (37B parameters) and DreamEngine (10.5B parameters), highlighting our framework’s effectiveness.

**Discussion and Connection to Methodology.** The strong performance of MENTOR — particularly its balanced multimodal generation and training efficiency — stems from its **autoregressive nature**

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

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

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 inputs and outputs. This direct alignment significantly enhances prompt following, ensuring generated images accurately reflect both text and visual guidance. The **two-stage training paradigm** is also critical for balanced multimodal control. It mitigates the common issue of over-reliance on one modality while ignoring others, resulting in significantly improved CP-PF scores compared to baselines. Notably, MENTOR achieves strong results despite using **relatively suboptimal components**. While other baselines rely on advanced models such as Qwen-2.5 and SD3, we use Flan-T5 as the encoder and LlamaGen as the generator — both of which greatly underperform stronger counterparts, as shown in Table 7 and Table 6. This highlights that our performance gains are driven by methodological strength, not from scaling up model capacity. While performance on out-of-distribution or fine-grained categories remains limited due to the suboptimal components, MENTOR maintains strong overall results by effectively integrate multimodal inputs for generation.

### 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 Stage 2 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 4, removing Stage 1 leads to the most 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 Stage 1 as critical for aligning multimodal inputs with output images. Without such alignment, 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 4. 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. Without these tasks, the model tends to over-rely on textual prompts, resulting in a regression towards a standard T2I model that neglects the reference image. These results highlight the importance of all Stage 2 tasks: 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.

### 3.4 Analysis

**Efficiency and Effectiveness: AR vs. Diffusion.** To evaluate the efficiency of our AR framework against diffusion-based approaches, 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 the same training data to ensure a fair comparison. Despite Kosmos-G employing



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	Photorealistic	Style Transfer	Imaginative	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	<b>0.55</b>	0.86	0.85	0.80	<b>0.84</b>	<b>0.47</b>

Table 4: Ablation studies on architecture design and multi-image training

Method	DreamBench++			DreamBench		
	CP	PF	CP-PF	DINOv1	CLIP-I	CLIP-T
<b>MENTOR</b>	0.555 $\pm$ 0.006	0.839 $\pm$ 0.002	0.466	70.853 $\pm$ 0.327	80.911 $\pm$ 0.053	29.071 $\pm$ 0.080
w. <i>Query-Variants</i>	0.421 $\pm$ 0.002	0.882 $\pm$ 0.000	0.371	54.518 $\pm$ 0.317	76.306 $\pm$ 0.114	30.792 $\pm$ 0.040
w. <i>Multi-image</i>	0.586 $\pm$ 0.006	0.829 $\pm$ 0.005	<b>0.486</b>	72.487 $\pm$ 0.147	81.857 $\pm$ 0.152	28.545 $\pm$ 0.043

a superior SD1.5 generator and Kosmos-1 as encoder, MENTOR, which utilizes a underperformed LlamaGen generator and a FlanT5 based encoder, demonstrated significantly better performance on DreamBench++ as shown in Table 3. It highlights the effectiveness of MENTOR in fostering efficient multimodal learning and strong multimodal conditional generation.

**Comparison of Proposed Architecture Variants.** As shown in Table 4, we compare two architectural variants of MENTOR: the MLP-based connector (MENTOR) and a Query-based connector for the multimodal encoder. It reveals that an MLP-based connector significantly outperforms the query-based variant in CP scores, especially for humans and objects. It suggests that after visual token compression, the query-based approach struggles to retain fine-grained visual details, which are crucial for generative fidelity, even when guided by textual queries. Nevertheless, due to effective token distillation, the Query-based variant facilitates training with multiple contextual images with minimal computational resources. Despite these differences, both variants exhibit competitive performance compared to other baselines in Table 3. This demonstrates that our simple and coherent architecture can effectively propagate visual features, whether or not perform aggressive visual token compression, highlighting the flexibility and robustness of our framework.

**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 4, 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.

**Image Reconstruction Fidelity.** To quantify visual detail preservation in our framework, we evaluate MENTOR on the Image Reconstruction Benchmark [Chen et al., 2025], 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 using pixel-space  $\ell_2$  distance, following pervious work [Chen et al., 2025]. As shown in Table 5, 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., 2025]—all of which couple LMMs with diffusion backbones. MENTOR achieves the best reconstruction quality, exceeding the second-best by 50%, even with a  $224 \times 224$  receptive field, while others varied from  $384 \times 384$  to  $512 \times 512$ . These gains confirm our model’s effectiveness at conditioning on—and faithfully reproducing—visual inputs.

Table 5: Image reconstruction performance.

Method	COCO ( $\downarrow$ )	JourneyDB ( $\downarrow$ )
SeedTokenizer	0.5102	0.5291
SEED-X	0.4317	0.4352
EMU2-Gen	0.3828	0.2869
DreamEngine	<u>0.2065</u>	<u>0.2052</u>
<b>MENTOR</b>	<b>0.1008</b>	<b>0.0867</b>

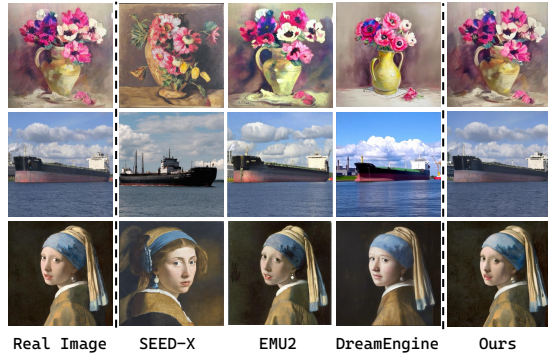


Figure 4: Qualitative study on Image Reconstruction.

**Versatility Across Different Multimodal Tasks.** To explore broader applicabilities of our framework, we evaluate its adaptability across diverse generation 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 1 show that the MENTOR produces coherent, high-quality outputs that adhere to the provided constraints without requiring any architectural modifications. While achieving performance in each specific domain would necessitate more specialized training and potentially more powerful multimodal encoder and generator components, these initial results underscore our framework’s versatility and its potential as an effective foundation for a variety of multimodal conditional image generation applications.

## 4 Related Work

### 4.1 Image Generation with Complex Multimodal Control

Researchers have developed image generation via diffusion models conditioned on multimodal inputs like canny edges [Zhang et al., 2023] and reference images [Zhao et al., 2024, Meng et al., 2022]. ControlNet [Zhang et al., 2023] uses auxiliary parameters, while Mix-of-Show [Gu et al., 2023] and FLUXSynID [Ismayilov et al., 2025] use LoRA modules for multi-concept control and identity preservation. DreamBooth [Ruiz et al., 2023] enable subject-specific fine-tuning but limit generalization. SuTI [Chen et al., 2023b] address it with scalable data and training. To enhance flexibility, recent work integrates LMMs with diffusion models by mapping LMM embedding into diffusion spaces [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., 2025] explore more complex multimodal prompt and fine-grained multimodal control. Yet, balancing guidance from diverse modalities remains a core challenge [Han et al., 2024, Ye et al., 2023, Mao et al., 2024]. EMMA [Han et al., 2024] 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. Nonetheless, these method often require substantial computational resources, and achieving efficient, robust, and scalable multimodal integration remains an open problem.

### 4.2 Autoregressive Multimodal Image Generation

Autoregressive models have driven progress in T2I generation, from DALL-E Ramesh et al. [2021] and Parti [Yu et al., 2022] to LlamaGen [Sun et al., 2024b] and GPT4O [OpenAI, 2024]. Recent work extends it to multimodal settings: Models like Chameleon [Team, 2024b], LWM Liu et al. [2024b], 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, 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] enhances Chameleon with omnipotent supervised fine-tuning for broader multimodal tasks. Nonetheless, these models often over-rely on visual inputs while ignoring textual prompts. 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 efficiency, scalability, and controllability of our method, establishing it as a efficient foundation for building versatile, fine-grained visual generation systems capable of handling complex multimodal prompts.

## References

- 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.
- BlackForestLabs. Announcing black forest labs, 2024. URL <https://blackforestlabs.ai/announcing-black-forest-labs/>.
- Bing Cao, Baoshuo Cai, Changqing Zhang, and Qinghua Hu. Dig2dig: Dig into diffusion information gains for image fusion, 2025. URL <https://arxiv.org/abs/2503.18627>.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 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, 2021.
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023a.
- Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think, 2025. URL <https://arxiv.org/abs/2502.20172>.
- Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning, 2023b. URL <https://arxiv.org/abs/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, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL <https://arxiv.org/abs/2210.11416>.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.
- 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. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv: 2309.11499*, 2023.
- Caption Emporium. midjourney-niji-1m-llavanext. <https://huggingface.co/datasets/CaptionEmporium/conceptual-captions-cc12m-llavanext>, 2024.
- Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2020. URL <https://api.semanticscholar.org/CorpusID:229297973>.
- 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. 2024.
- Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. ISSN 0262-8856. doi: <https://doi.org/10.1016/j.imavis.2024.105171>. URL <https://www.sciencedirect.com/science/article/pii/S0262885624002762>.

- Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022. URL <https://arxiv.org/abs/2203.13131>.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL <https://arxiv.org/abs/2208.01618>.
- Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.
- 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.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2024.
- Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models, 2023. URL <https://arxiv.org/abs/2305.18292>.
- Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts, 2024. URL <https://arxiv.org/abs/2406.09162>.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- Raul Ismayilov, Dzemila Sero, and Luuk Spreeuwers. Fluxsynid: A framework for identity-controlled synthetic face generation with document and live images, 2025. URL <https://arxiv.org/abs/2505.07530>.
- 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. Segment anything. *ICCV*, 2023.
- Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023.
- Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, 2023a. URL <https://arxiv.org/abs/2305.14720>.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597, 2023b.
- Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. UNIMO-G: Unified image generation through multimodal conditional diffusion. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6173–6188, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.335. URL <https://aclanthology.org/2024.acl-long.335/>.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024a.

- Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2024b. URL <https://arxiv.org/abs/2402.08268>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485, 2023.
- Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action, 2023.
- Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom++: Representing images as real-word for real-time customization, 2024. URL <https://arxiv.org/abs/2408.09744>.
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. URL <https://arxiv.org/abs/2108.01073>.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. URL <https://arxiv.org/abs/2112.10741>.
- OpenAI. hello-gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>.
- Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhui Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models, 2024. URL <https://arxiv.org/abs/2310.02992>.
- Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation, 2025. URL <https://arxiv.org/abs/2406.16855>.
- 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, 2023. URL <https://arxiv.org/abs/2307.01952>.
- 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.
- 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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL <https://api.semanticscholar.org/CorpusID:231591445>.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL <https://arxiv.org/abs/2102.12092>.
- 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.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022a.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022b. URL <https://arxiv.org/abs/2112.10752>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022c. URL <https://arxiv.org/abs/2112.10752>.

- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. URL <https://arxiv.org/abs/2208.12242>.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.
- Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: LLaMA for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024a.
- Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024b. URL <https://arxiv.org/abs/2406.06525>.
- Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality, 2023.
- Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiyang Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024c.
- Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. URL <https://arxiv.org/abs/2411.15098>.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024a.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024b.
- Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Xiaoming Wei, and Enhua Wu. Image editing with diffusion models: A survey, 2025. URL <https://arxiv.org/abs/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. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024a. URL <https://arxiv.org/abs/2409.12191>.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024b. URL <https://arxiv.org/abs/2409.18869>.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024a. URL <https://arxiv.org/abs/2410.13848>.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024b.
- 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. Vila-u: a unified foundation model integrating visual understanding and generation, 2024c. URL <https://arxiv.org/abs/2409.04429>.

- Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. URL <https://arxiv.org/abs/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. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding, 2025. URL <https://arxiv.org/abs/2411.17762>.
- Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.
- Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.
- 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. Anygpt: Unified multimodal llm with discrete sequence modeling. *ArXiv*, abs/2402.12226, 2024. URL <https://api.semanticscholar.org/CorpusID:267750101>.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.
- Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *ArXiv preprint*, abs/2309.07915, 2023.
- Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale, 2024. URL <https://arxiv.org/abs/2407.05282>.
- Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. *arXiv preprint arXiv:2406.18583*, 2024.



## A Technical Appendices and Supplementary Material

This Supplementary Material is organized as follows.

- In Appendix B, we provide the training details of MENTOR, including initialization (Appendix B.1), training procedures (Appendix B.2), and multi-image training strategy (Appendix B.3).
- 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 discuss the current limitations of our method, such as its dependence on autoregressive generators, generation fidelity, and safety considerations.

## 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  $224 \times 224$ , 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.

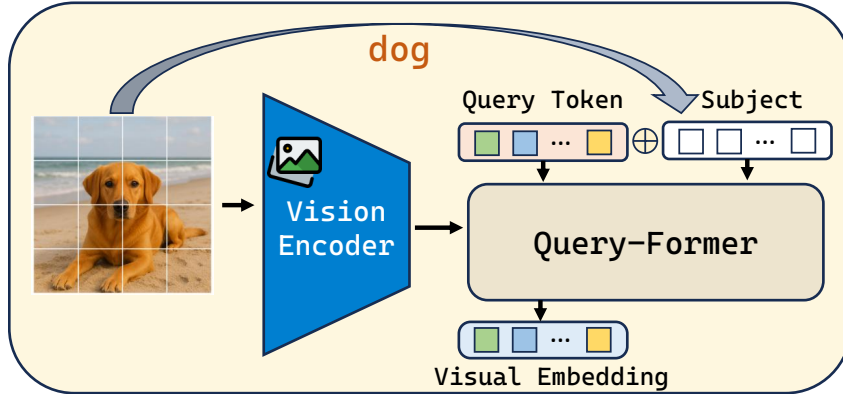


Figure 5: Overview of text-guided visual distillation using the Query-based variant of MENTOR.

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., 2024b] 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

\*<https://github.com/FoundationVision/LlamaGen/issues/54>

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  $5 \times 10^{-4}$ , 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  $1 \times 10^{-4}$ , 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 **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 6, demonstrate that multi-image training significantly enhances the model’s capability to preserve detailed visual information in complex multimodal contexts.

## C Text-to-Image Generation Evaluation

We evaluate the performance of our model on text-to-image (T2I) generation using the MS-COCO [Lin et al., 2014] and GenEval [Ghosh et al., 2024] benchmarks. Results are reported in Table 7 and 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 6: Qualitative assessment demonstrating improved preservation of visual details by MENTOR following multi-image training.

## D Data Construction and Formation

**Data Formation** Table 8 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].

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
<i>Autoregressive</i>	Chameleon Team [2024a]	-	-	-	-	-	-	0.39
	LWM Liu et al. [2024a]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	LlamaGen Sun et al. [2024a]	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	0.38	0.18	0.67	0.08	0.13	0.38
<i>Diffusion</i>	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- $\alpha$ 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., 2025]	1.00	0.94	0.64	0.81	0.27	0.49	0.69
	SDv3.5 Large Esser et al. [2024]	0.98	0.89	0.73	0.83	0.34	0.47	0.71

Table 7: Zero-shot FID scores on the MS-COCO benchmark. Lower is better.

Model	FID ↓
<i>T2I Models</i>	
GLIDE [Nichol et al., 2022]	12.24
Make-A-Scene [Gafni et al., 2022]	11.84
DALL-E 2 [Ramesh et al., 2022]	10.39
SD v1.5 [Rombach et al., 2022c]	9.34
Imagen [Saharia et al., 2022]	7.27
<i>CLIP-Aligned VL2I Models</i>	
GILL [Koh et al., 2023]	12.20
Emu [Sun et al., 2023]	11.66
KOSMOS-G [Pan et al., 2024]	10.99
MENTOR	19.92

Table 8: 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

We further incorporate 400k subject-driven image generation samples from Subject200k [Tan 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 7. 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:

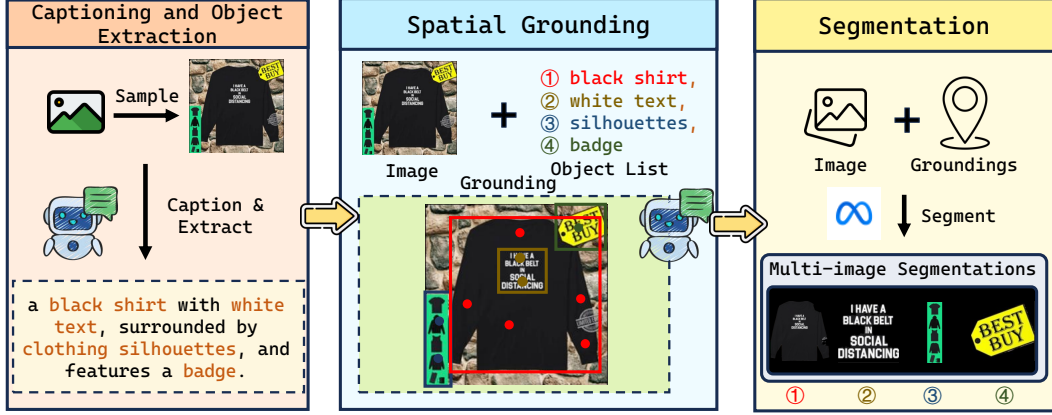


Figure 7: Illustration of the automatic data generation pipeline.

- **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 keypoints. 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 diverse 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** ( $150 \times 9$ ), representing a substantial scale-up over the original DreamBench (30 subjects  $\times$  25 prompts). Relative to DreamBench, the dataset is **5 $\times$  larger in subjects** and **54 $\times$  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.

**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:

- **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., 2025]: 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-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 models 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 8 and 9, 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.



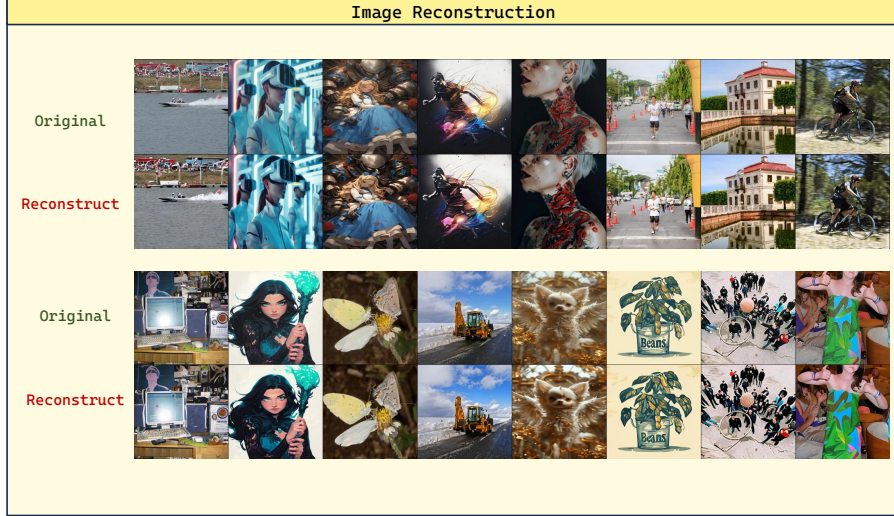


Figure 8: Qualitative comparison of image reconstruction results using MENTOR.

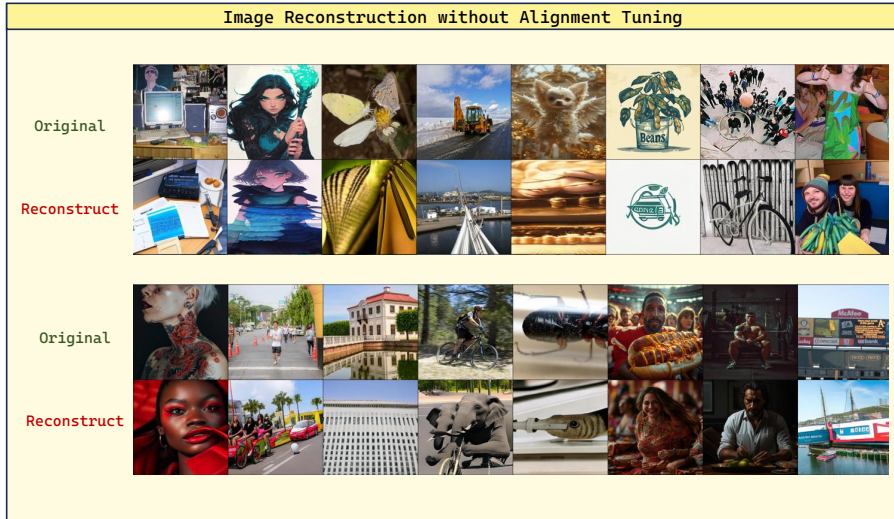


Figure 9: Image reconstruction results of MENTOR *without* alignment tuning.

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 9, 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 10, 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.

## F.3 Multi-Image Generation

We evaluate MENTOR on multi-image generation tasks using the X2I dataset [Xiao et al., 2024]. As shown in Figure 11, the model is capable of generating visually consistent outputs conditioned the

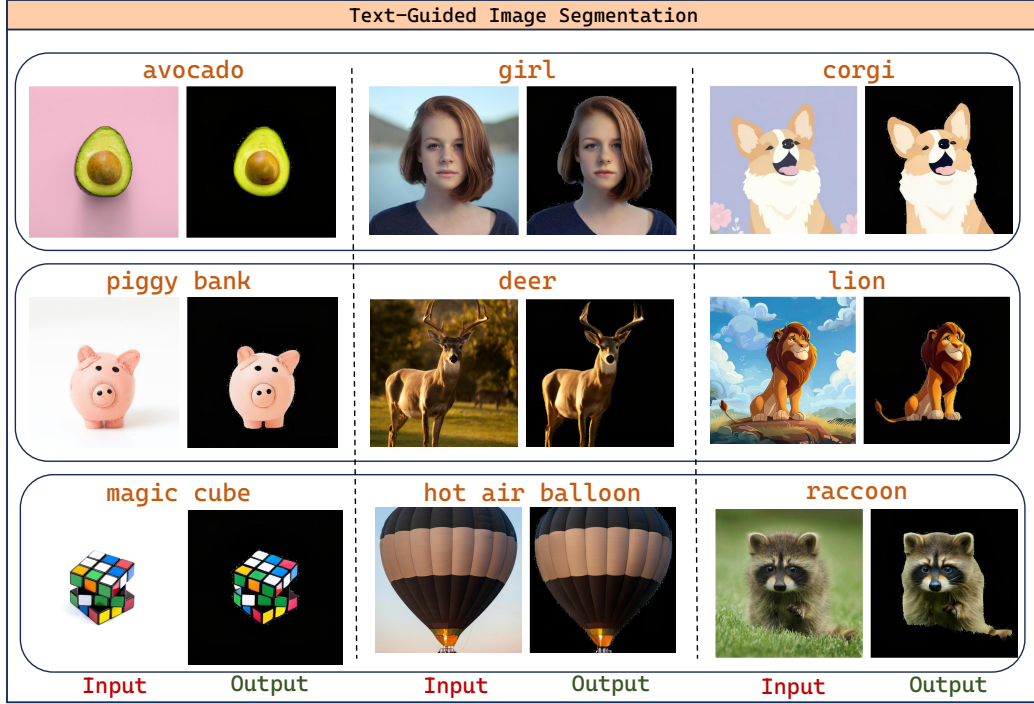


Figure 10: Qualitative results of text-guided image segmentation using MENTOR.

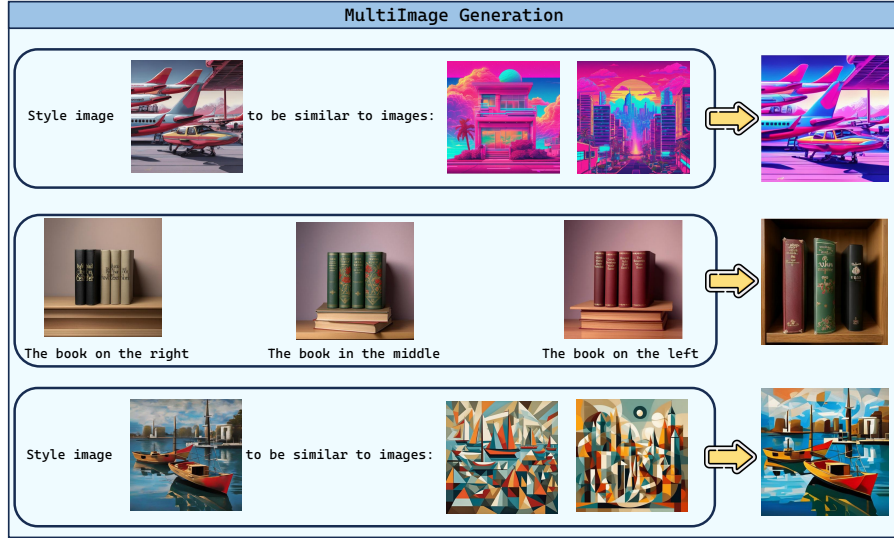


Figure 11: Qualitative results for multi-image generation.

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., 2024]. As illustrated in Figure 12, MENTOR learns to synthesize images that follow the stylistic patterns demonstrated in the in-context



Figure 12: Qualitative examples from multimodal in-context image generation. The model adapts to patterns in the visual 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.

**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., 2024] dataset for two additional epochs using a reduced learning rate of  $5 \times 10^{-5}$ . 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., 2024], which features sequences of input-output pairs for in-context generalization. We use a learning rate of  $5 \times 10^{-5}$  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 prompting 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 Limitations

Our work presents a promising alternative to diffusion-based methods for multimodal-conditioned image generation. As such, our focus is on evaluating performance under comparable conditions—i.e., with similar model capacities and training paradigms. However, the effectiveness of our approach is currently constrained by the limitations of available autoregressive (AR) backbone models. Due to the lack of high-performance AR generators, MENTOR exhibits shortcomings in several aspects of 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. 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. These initial findings nonetheless highlight the framework’s potential as a unified and efficient foundation for conditional multimodal image generation.