MANZANO: A SIMPLE AND SCALABLE UNIFIED MULTIMODAL MODEL WITH A HYBRID VISION TOKENIZER

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal Large Language Models (LLMs) that can both understand and generate visual content hold immense potential. However, existing open-source models often suffer from a performance trade-off between these capabilities. We present Manzano, a simple and scalable unified framework that substantially reduces this tension by coupling a hybrid image tokenizer with a well-curated training recipe. A single shared vision encoder feeds two lightweight adapters that produce continuous embeddings for image-to-text understanding and discrete tokens for text-to-image generation within a common semantic space. A unified autoregressive LLM predicts high-level semantics in the form of text and image tokens, with an auxiliary diffusion decoder subsequently translating the image tokens into pixels. The architecture, together with a unified training recipe over understanding and generation data, enables scalable joint learning of both capabilities. Manzano achieves state-of-the-art results among unified models, and is competitive with specialist models, particularly on text-rich evaluation. Our studies show minimal task conflicts and consistent gains from scaling model size, validating our design choice of a hybrid tokenizer.

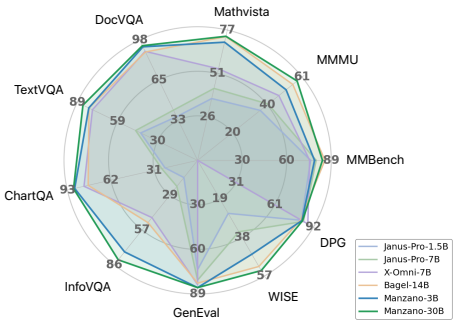
1 INTRODUCTION

Unified multimodal models (OpenAI, 2025; Deng et al., 2025; Chen et al., 2025d; Wu et al., 2025a; Zhou et al., 2024), which integrate both understanding and generation capabilities, have become increasingly prominent within the research community. The appeal of this paradigm stems from the discovery that integrating these domains unlocks emergent capabilities (OpenAI, 2025; Deng et al., 2025) in generation, such as complex world reasoning, multimodal instruction following, and iterative visual editing. Yet, in practice, adding generation often degrades understanding. Existing unified models (Deng et al., 2025; Fan et al., 2025; Liang et al., 2024; Chen et al., 2025d) consistently lag far behind their understanding-only counterparts (Seed, 2025; Bai et al., 2025; Zhang et al., 2024a), especially on text-rich benchmarks (Mathew et al., 2021; Masry et al., 2022).

A key reason for this gap is the conflicting nature of visual tokenization. Auto-regressive generation usually prefers discrete image tokens (Chameleon, 2024; Wu et al., 2024; Ma et al., 2025) while understanding typically benefits from continuous embeddings. Many models adopt a dual-tokenizer strategy (Wu et al., 2025b; Chen et al., 2025d; Fan et al., 2025; Tong et al., 2024b), using a semantic encoder for rich, continuous features while a separate quantized tokenizer like VQ-VAE (Van Den Oord et al., 2017) handles generation. However, this forces the language model to process two different image token types, one from high-level semantic space versus one from low-level spatial space, creating a significant task conflict. While some solutions like Mixture-of-Transformers (MoT) (Liang et al., 2024; Deng et al., 2025) can mitigate this by dedicating separate pathways for each task, they are parameter-inefficient and are often incompatible with modern Mixture-of-Experts (MoE) (Fedus et al., 2022; Lepikhin et al., 2021) architectures. An alternative line of work bypasses this conflict by freezing a pre-trained multimodal LLM and connecting it to a diffusion decoder (Pan et al., 2025; Wu et al., 2025a;c). While this preserves the understanding capability, it decouples generation, losing potential mutual benefits and limiting potential gains for generation from scaling the multimodal LLM.

To overcome the above challenges, we propose **Manzano**, a simple unified model that harmonizes the representations for understanding and generation. Manzanoemploys a unified shared visual encoder

(a) **Qualitative text-to-image generation** on challenging prompts. Manzano handles counterintuitive, physics-defying prompts (e.g., 'The bird is flying below the elephant') comparably to GPT-4o (Islam & Moushi, 2024) and Nano Banana (DeepMind, 2025).



(b) **Quantitative comparisons** on popular understanding and generation benchmarks. Manzano 3B and 30B models achieve superior or competitive performance compared to other SOTA unified multimodal LLMs.

Figure 1: Comparison of qualitative and quantitative results for Manzano model.

with two lightweight and specialized adapters: a *continuous* adapter for understanding tasks and a *discrete* adapter for generation. Because two adaptors originate from the same encoder, it yields hybrid representations from a homogeneous source, significantly mitigating task conflict in the LLM. We first pre-train the hybrid tokenizer with a small LLM decoder to pre-align the image features with the LLM feature space. Then the autoregressive multimodal LLMs are jointly trained on a mixture of pure text, image understanding, and image generation data. Finally, we leverage a diffusion image decoder (Peebles & Xie, 2023; Chen et al., 2025a) to render pixels by taking the generated image tokens as conditioning.

We train the unified multimodal LLM with a joint recipe to learn image understanding and generation simultaneously. This training consists of three stages: a pre-training stage on a large-scale corpus of text-only, interleaved image-text, image-to-text (IT), and text-to-image (TI) data; a continued pre-training stage on higher-quality IT and TI data; and a supervised fine-tuning (SFT) stage on curated text, IT, and TI instruction data to enhance instruction following capability and improve both understanding and generation tasks.

We demonstrate that Manzano achieves state-of-the-art performance on both understanding and generation tasks. As shown in Fig. 1a and 1b, our 3B model, despite its smaller LLM size, achieves competitive generation performance compared to other unified multimodal LLMs. Simultaneously, it delivers significantly better understanding performance, especially on text-rich benchmarks that demand precise perceptual capabilities. Our ablations on the training recipes also indicate minimal cross-task conflict under joint training (Fig. 3). These findings suggest that the architecture and the training recipe effectively mitigate the conflict between understanding and generation, even in a compact model.

Facilitated by the simplicity of the architecture and the joint training recipe, we further investigate the *scaling* behavior of Manzano. Our scaling studies in Sec. 4.3 show substantial improvements across both understanding and generation benchmarks when scaling the LLM decoder (from 300M to 30B). In addition, enlarging the diffusion decoder also leads to significant gains in image structural integrity, as validated by large-scale human evaluations.

2 Related Work

2.1 MLLMs for Image Understanding

Recent advances in Multimodal Large Language Models (MLLMs) have led to a widely adopted architectural pattern that links a vision encoder with a language model through a trainable interface. Typical vision encoders include CLIP (Radford et al., 2021a), SigLIP (Zhai et al., 2023), and the recent InternViT (Chen et al., 2024c). Early works experimented with elaborate connector

designs—for example, Flamingo (Alayrac et al., 2022) incorporates gated cross-attention layers to inject image features into the LLM, while BLIP-2 (Li et al., 2023b) introduces the Q-Former to better align visual and textual representations. A notable departure from these complex strategies is LLaVA (Liu et al., 2023; 2024a), which demonstrates that a lightweight Multi-Layer Perceptron (MLP) projection can effectively serve as the connector. This simplicity has since become the blueprint for many follow-up systems, such as the MM1 series (McKinzie et al., 2024; Zhang et al., 2024a), the InternVL family (Chen et al., 2024c; Zhu et al., 2025; Wang et al., 2025a), and the Qwen-VL models (team, 2024; Bai et al., 2025), which further improve performance by scaling up both data and backbone models. However, these MLLMs are primarily designed for understanding tasks and lack the capability to generate high-quality images, which limits their applicability in tasks that require bidirectional visual–text reasoning and creation. Despite being limited to understanding tasks, MLLMs still provide valuable strengths—their training recipes and scaling strategies are much more mature than those of current unified multimodal models, which we discuss next.

2.2 Unified Multimodal Models

The integration of image understanding and generation within a single, unified multimodal LLM is becoming prominent. GPT-4o (OpenAI, 2025) demonstrates embedding image generation capabilities directly into an autoregressive LLM, which unlocks emergent abilities, such as stronger instruction following, improved text rendering, multi-turn visual editing, and sophisticated world knowledge reasoning. Existing unified models can be broadly categorized into three architectural paradigms. First, the unified autoregressive (AR) approach (OpenAI, 2025; Chen et al., 2025d;c;b; Han et al., 2025b; Tian et al., 2024; Chameleon, 2024; Tong et al., 2024b; Fan et al., 2025; Ma et al., 2025; Han et al., 2025a; Geng et al., 2025; Wu et al., 2025d; 2024) converts images into sequences of discrete or continuous tokens, enabling LLM to jointly model both image and text sequences in an autoregressive manner. Second, the decoupled LLM-diffusion approach (Pan et al., 2025; Wu et al., 2025a;c) employs a largely frozen LLM for semantic understanding and contextual reasoning, while delegating image synthesis to a separate diffusion decoder. In this design, the LLM itself does not possess native image generation capability. Third, the hybrid AR-diffusion approach (Deng et al., 2025; Zhou et al., 2024; Liang et al., 2024) integrates both paradigms within a single transformer, using autoregressive decoding for text and an embedded diffusion process for images. Our model is most closely aligned with the first, autoregressive paradigm. However, instead of employing separate tokenizers (Deng et al., 2025; Chen et al., 2025d; Fan et al., 2025; Wu et al., 2025c) for understanding and generation, we introduce a unified semantic tokenizer to produce both continuous features for understanding tasks and quantized features for generation tasks. This hybrid tokenizer strategy substantially mitigates the task conflict that commonly arises. Moreover, while our LLM backbone follows the autoregressive design, we augment it with a diffusion decoder for image synthesis, enabling high-fidelity generation guided by the semantic representations generated by the LLM.

2.3 Diffusion Models for Image Generation

Diffusion-based generative models (Song et al., 2020; Ho et al., 2020; Dhariwal & Nichol, 2021) have become one of the most prominent approaches for high-fidelity image synthesis. These models gradually refine Gaussian noise into realistic images through a learned denoising process. Latent diffusion methods (Rombach et al., 2022; Podell et al., 2023) enhance computational efficiency by conducting generation in the latent space of a pre-trained variational autoencoder (VAE) (Kingma & Welling, 2022), reducing memory and compute requirements while preserving visual quality. More recently, flow matching approaches (Liu et al., 2022; Ma et al., 2024; Tong et al., 2023) have been introduced to connect source and target distributions via simplified continuous trajectories, leading to further gains in synthesis performance (Esser et al., 2024). In parallel, architectural advances such as Diffusion Transformers (DiTs) (Peebles & Xie, 2023; Chen et al., 2024a; Esser et al., 2024; Labs, 2024; Chen et al., 2025a) have demonstrated strong scalability and quality improvements, echoing the success of transformer-based designs in natural language processing.

Building on these developments, our diffusion decoder integrates the strengths of the field by employing a DiT in the latent domain with a conditional flow matching objective. Unlike conventional text-to-image diffusion models (Ramesh et al., 2022; Saharia et al., 2022) conditioned on semantic embeddings from pre-trained text encoders such as CLIP (Radford et al., 2021a), our method leverages visual token embeddings generated by LLM as conditioning signals.

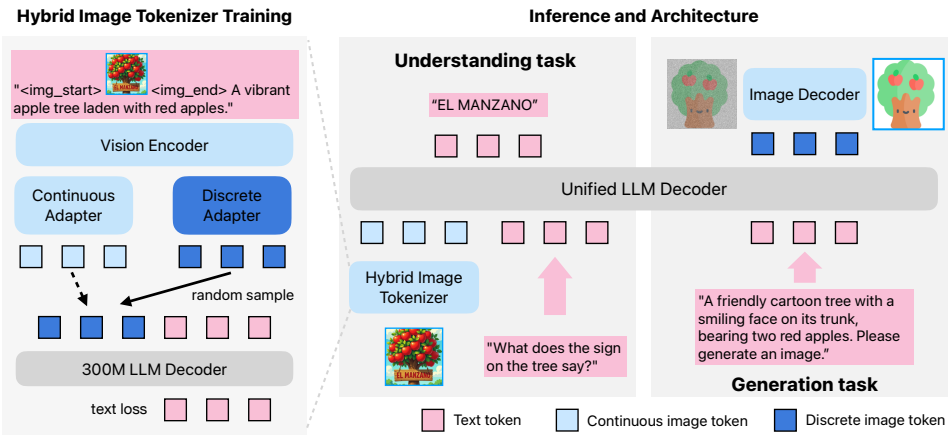


Figure 2: **Our hybrid tokenizer workflow**. **(Left)**: The tokenizer produces two distinct but homogeneous feature streams through separate adapters. During training, one adapter output is randomly sampled and passed to a small LLM decoder for alignment. **(Right)**: Once the tokenizer is trained, the right panel illustrates how these two feature types are applied to understanding and generation tasks.

3 MODEL

Manzano is a multimodal large language model (MLLM) that unifies understanding and generation tasks using the auto-regressive (AR) approach. The architecture comprises three components: (i) a *hybrid vision tokenizer* that produces both continuous and discrete visual representations; (ii) an *LLM decoder* that accepts text tokens and/or continuous image embeddings and auto-regressively predicts the next discrete image or text tokens from a joint vocabulary; and (iii) an *image decoder* that renders image pixels from predicted image tokens (see Figure 2 for the framework).

3.1 DESIGN CHOICES

Unified hybrid representation. The hybrid image tokenizer encodes images into continuous tokens for understanding (I2T), and discrete tokens for generation (T2I), while sharing the same visual encoder: (i) *Continuous for I2T*: Manzano utilizes continuous embeddings for I2T tasks, a strategy widely adopted in popular visual understanding models (team, 2024; Seed, 2025), which has proven superior performance, especially on text-rich tasks that require more visual details (e.g., DocVQA, ChartQA, and InfoVQA). Our ablation (Table 1) also shows discrete tokens underperform on understanding tasks, which reflects the weak understanding results reported for some pure-discrete unified models (Chameleon, 2024; Wang et al., 2024). (ii) *Discrete for T2I*: Representing images as discrete code indices lets the LLM use the same AR next-token learning strategy as text, simplifying the generation pipeline and scaling behavior. (iii) *Shared unified semantic space*: Both branches originate from the same encoder backbone; thus, continuous and discrete tokens inhabit a common semantic space, which reduces potential task conflict. The LLM decoder focuses on regressing high-level semantics (text and image tokens), while the diffusion decoder is responsible for rendering high-fidelity details in pixel space. Many existing unified models rely on separate tokenizers for understanding and generation (Chen et al., 2025d; Deng et al., 2025) — for instance, using a CLIP tokenizer for understanding tasks and a VAE tokenizer for generation. Although this strategy preserves more image spatial details, it exacerbates the task conflict within the subsequent LLM. Some studies (Chen et al., 2025b;c) find that a dedicated generation tokenizer is not as compatible with LLM as the semantic tokenizer. Thus, our hybrid unified image tokenizer employs a single image encoder for both understanding and generation tasks.

Simplicity and scalability. Our design keeps the training losses standard and components cleanly decoupled, which simplifies unification and scaling for the unified MLLM in these aspects: (i) *Unified AR objective*: Our unified LLM decoder uses a single AR objective for text-only, I2T, and T2I tasks without additional auxiliary losses or per-task heads. (ii) *Decoupled components*: The clear split between semantic prediction (LLM decoder) and detail generation (image decoder) supports

independent scaling of the base LLM and the image decoder. (iii) *Practical scaling*: Our approach readily leverages mature, scalable training pipelines from LLM/MLLM and diffusion decoders. By contrast, prior works (e.g., Transfusion (Zhou et al., 2024) and Bagel (Deng et al., 2025)) explored incorporating auto-regressive text prediction and a diffusion image generation process for image generation in one LLM, but leave large-scale scaling under-explored. Our decoupled design facilitates scaling the LLM decoder to 30B and diffusion decoder to 3B, yielding promising scaling behavior (Sec. 4.3).

3.2 Architecture

Hybrid Image Tokenizer. Our tokenizer comprises three components: (i) a standard vision transformer (ViT) (Dosovitskiy et al., 2020) as the vision backbone; (ii) a continuous adapter, which first applies a 3×3 Spatial-to-Channel (STC) layer to reduce the number of spatial tokens by a factor of 9 (e.g., from $42 \times 42 \times 1024$ to $14 \times 14 \times 9216$) and then uses an MLP to project each feature into the LLM feature dimension (e.g., 2048); and (iii) a discrete adapter, which also starts with the STC compression step but further quantizes the features using finite scalar quantization (FSQ) (Mentzer et al., 2023) — chosen for its simplicity and scalability to large codebooks (64K in our experiments) — before applying an MLP projection into the LLM feature dimension.

Unified LLM. We connect our hybrid image tokenizer to a standard text LLM decoder for unified training on a mixture of datasets containing text, understanding, and generation data. For the language backbone, we leverage internal pre-trained LLMs.

Image Decoder. We train an image decoder on top of a pre-trained hybrid image tokenizer to reconstruct images in pixel space from discrete image tokens. Given an input image, the hybrid tokenizer first encodes it into a latent representation, which serves as the conditioning input for a flow-matching pipeline (Lipman et al., 2022) that transports Gaussian noise into realistic images. For the decoder backbone, we adopt the DiT-Air architecture (Chen et al., 2025a), which employs a layer-wise parameter-sharing strategy that reduces the size of the standard MMDiT model (Esser et al., 2024) by approximately 66% while maintaining comparable performance. We provide three decoder configurations with the parameter size of 0.9B, 1.75B, and 3.52B, supporting a range of output canvas resolutions from 256 to 2048 pixels.

Inference Pipeline. The inference pipeline for both understanding and generation tasks is shown in Fig. 2 (right). For understanding tasks, Manzano uses the hybrid image tokenizer to extract continuous features. These features, along with text features, are then fed into the unified LLM decoder to predict the final answer. For generation tasks, Manzano takes a text input and predicts a sequence of image tokens. The image decoder then renders these tokens into image pixels.

4 Experiments

In this section, we first introduce the evaluation setup for understanding and generation capabilities (Sec. 4.1). We then study the cross-token interplay in our model (Sec. 4.2). After that, we analyze the scaling behavior by varying model sizes (Sec. 4.3). Finally, we compare our model against state-of-the-art models, including both specialist and unified models. The training details, including the data mixture for understanding and generation and the unified training recipe, can be found in Sec. A.

4.1 Evaluation

We evaluate our models on image understanding and generation capabilities on popular benchmarks.

Understanding. We adopt the three categories of benchmarks for multimodal understanding: (i) *General VQA*: SeedBench (Li et al., 2023a), RealWorldQA (Zhang et al., 2024b), and MMBench (Liu et al., 2024b). (ii) *Knowledge & Reasoning*: AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2023), and MathVista (Lu et al., 2023). (iii) *Text-rich Document & Chart Understanding*: ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), and OCRBench (Liu et al., 2024c).

Tokenizer Paradigm	Understanding Tasks			Generation Tasks		
	General	Knowledge	Text-Rich	GenEval	DPG	WISE
Pure-Discrete	63.3	62.2	62.3	77	80.9	35
Dual-Encoder	63.8	63.6	72.0	65	66.3	17
Hybrid Tokenizer	64.9	66.5	73.3	77	79.9	35

Table 1: **Tokenizer strategy ablation.** Tokenizers are evaluated with a 1B unified LLM model. The full list of evaluation tasks for understanding can be referred to Sec. 4.4.1. The hybrid tokenizer outperforms the other two tokenizer paradigms.

Generation. We use both automated and human evaluations: (i) *Automated Evaluation*: The automated benchmarks include GenEval (Ghosh et al., 2023) and DPGBench (Hu et al., 2024) for prompt following generation, and WISE (Niu et al., 2025) for World Knowledge-Informed generation. (ii) *Human Evaluation*: We curate a comprehensive evaluation set comprising 800 challenging prompts, subsampled from established academic benchmarks (Wiles et al., 2025; Yu et al., 2022) and from widely used community evaluation platforms. The generated outputs are assessed by in-house human raters on three dimensions: structural integrity, instruction following, and aesthetic quality. For each dimension, raters assign one of three grades: major issues, minor issues, or no issues, and are quantized to scores afterwards. To mitigate bias, entity information is masked, and the sample order is randomized. Each sample is independently rated by three raters, and the final scores are obtained by averaging across raters to reduce variability.

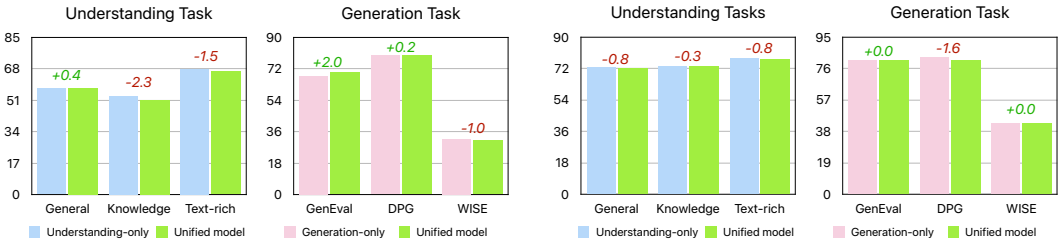
4.2 Understanding-Generation Interplay

In this section, we study the task conflict along two axes: (i) *tokenizer strategy* (pure-discrete vs. dual-encoder vs. our hybrid); (ii) *task mixing* (unified vs. single-task). For simplicity, we skip the continued pre-training stage in the unified LLM training for these ablations.

Tokenizer Strategy. We construct two baselines to compare our unified hybrid tokenizer strategy: (i) *Pure-discrete.* Prior works (Chameleon, 2024; Wang et al., 2024; Wu et al., 2024) train a quantized semantic vision tokenizer using various quantization techniques (Mentzer et al., 2023; Van Den Oord et al., 2017) and then use an LLM to predict the next text and image tokens. To mimic these methods in our setting, we replace the understanding inputs for LLM with discrete features from our hybrid tokenizer, so the LLM uses the same discrete tokens for both understanding and generation. To isolate the effect of quantization on understanding, we use the same weights for the vision encoder and the discrete adapter from our hybrid tokenizer. (ii) *Dual-encoder.* Another popular models (Chen et al., 2025d; Deng et al., 2025) uses a dual-encoder strategy to preserve detailed features by a semantic encoder for understanding and a VAE-style encoder for generation, effectively mitigating the degradation of understanding. We reproduce this baseline by replacing the discrete tokens from our hybrid tokenizer with those generated by an internal reproduction of MagViT-2 (Yu et al., 2023), an autoencoder-style tokenizer. This MagViT-2 tokenizer uses FSQ (Mentzer et al., 2023) with a 64K codebook and a spatial compression ratio of 8. For generation tasks, we resize images to 128x128 pixels instead of the original 256x256. This reduced the number of tokens per image to 256, which we found improved the model's instruction-following capabilities on benchmarks.

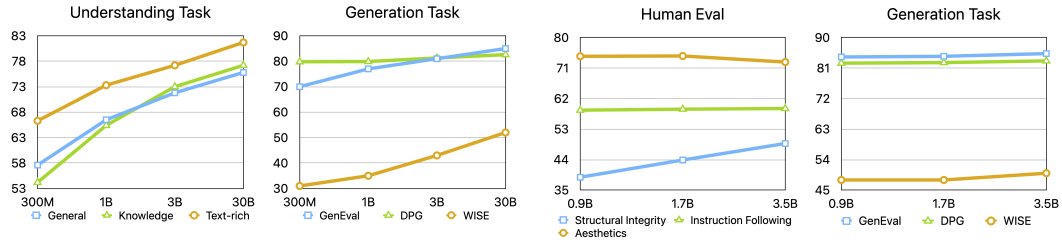
Table 1 shows the results on both image understanding and generation tasks. Our hybrid tokenizer paradigm shows the least task conflict and outperforms both pure-discrete and dual-encoder baselines on all tasks. Specifically, the pure-discrete baseline leads to a significant drop in understanding performance–especially on text-rich benchmarks, due to information loss from quantization. While the dual-encoder baseline mitigates some of this degradation, it still consistently underperforms our hybrid tokenizer on all understanding tasks–especially on knowledge benchmarks, which rely heavily on the LLM's reasoning abilities. This suggests that the conflict between heterogeneous visual tokens resides within the LLM.

Unified vs. Single-task. To quantify the task conflict in our hybrid tokenizer paradigm, we compare our unified model with baselines trained exclusively for understanding or generation. For the understanding-only baseline, we remove all text-to-image data from both the pre-training and SFT stages. We reduce the training steps to ensure it is exposed to the same number of text and image understanding tokens as our unified model. Similarly, for the generation-only baseline, we remove



(a) 300M Model (b) 3B Model

Figure 3: **Unified *vs*. Single-task study.** Our unified model exhibits a slight regression compared with the understanding-only model on understanding task; however, this effect becomes negligible at the 3B scale, where the gap is less than 1.0. For generation, the unified model shows a decline on only one benchmark compared with the generation-only model.



(a) LLM Decoder Scaling (b) Image Decoder Scaling

Figure 4: **Model scaling behavior of Manzano**. (a) Scaling the LLM decoder yields monotonic improvements across both understanding and generation benchmarks. (b) Scaling the image decoder enhances structural integrity while maintaining stable quantitative benchmarks. A drop in aesthetic quality is observed, which we leave for more in-depth study in future work.

the understanding data and keep only the text-only and text-to-image data, while also reducing the training steps. We conduct this ablation study with a 300M and a 3B LLM decoder. The results, plotted in Fig. 3a and 3b, show that the unified LLM trained with our hybrid tokenizer performs on par with the dedicated, single-task models on nearly all tasks, even at a compact size like 300M. This demonstrates that our unified hybrid tokenizer paradigm successfully unifies visual perception and generation without a performance trade-off.

4.3 Model Scaling Behavior

Facilitated by the decoupled design of LLM Decoder and Image Decoder, we explore the model scaling behavior along two dimensions: *LLM Decoder* and *Image Decoder*. Similar to Sec. 4.2, we skip the continued pre-train stage in the unified LLM training for the scaling experiments.

Scaling LLM Decoder. We vary only the LLM Decoder size (300M, 1B, 3B, and 30B) while keeping the image decoder (0.9B), data mixtures, and training hyperparameters fixed[1]. Fig. 4a shows monotonic gains across all understanding (General / Knowledge / Text-Rich) and generation (GenEval / DPG / WISE) metrics as the LLM decoder scales. Compared to 300M, our 3B Manzano model improves significantly by +14.2 (General), +18.8 (Knowledge), +10.9 (Text-rich), +11.0 (GenEval), +1.48 (DPG), +12.0 (WISE). Further scaling to 30B yields smaller but consistent gains over 3B. Fig. 7 shows the qualitative examples for image generation. We can see that the generation capabilities, including instruction-following, text-rendering, and overall image quality, are improved consistently across different LLM scales. The results support the simple yet effective design for Manzano: LLM decoder captures high-level semantics, and scaling it benefits both understanding and generation.

Scaling Image Decoder. We evaluate the performance of image decoders of varying sizes built on top of a 3B LLM decoder. Figure 4b shows that, in human evaluations, structural integrity improves

[1]We pre-train 30B LLM Decoder on roughly half the tokens compared to other model sizes.

Model	General Benchmarks			Knowledge Benchmarks				Text-rich Benchmarks				
	SEED[I]	RealWorldQA	MMBench (dev-en)	AI2D (test)	SQA (test)	MMMU (val)	MathV (testmini)	ChartQA (test)	TextVQA (val)	DocVQA (test)	InfoVQA (test)	OCRBench (test)
3B Specialist Model												
BLIP-3-4B (Xue et al., 2024)	72.2	60.5	–	–	88.3	41.1	39.6	–	71.0	–	–	–
Phi-3-Vision-4B (Abdin et al., 2024)	71.8	59.4	78.6	76.7	90.8	40.4	44.5	81.4	70.1	83.3	49.0	63.7
MM1.5-3B (Zhang et al., 2024a)	72.4	56.9	72.4	65.7	85.8	37.1	44.4	74.2	76.5	87.7	58.5	65.7
InternVL2.5-2B (Chen et al., 2024b)	–	60.1	77.2	74.9	–	43.6	51.3	79.2	74.3	88.7	60.9	80.4
InternVL2.5-4B (Chen et al., 2024b)	–	64.3	78.7	81.4	–	52.3	60.5	84.0	76.8	91.6	72.1	82.8
InternVL3.5-4B (Wang et al., 2025b)	–	66.3	–	82.6	–	66.6	77.1	86.0	77.9	92.4	78.0	82.2
Qwen2.5VL-3B (Bai et al., 2025)	–	65.4	76.4	81.6	–	53.1	62.3	84.0	79.3	93.9	77.1	79.7
30B Specialist Model												
LLaVA-NeXT-34B (Liu et al., 2024a)	75.9	–	–	–	81.8	51.1	46.5	–	69.5	–	–	–
Cambrian-34B (Tong et al., 2024a)	75.3	67.8	–	79.7	85.6	49.7	53.2	75.6	76.7	75.5	–	60.0
MM1.5-30B (Zhang et al., 2024a)	75.0	69.0	–	77.2	91.9	47.4	55.6	83.6	79.2	91.4	67.3	65.8
InternVL2.5-26B (Chen et al., 2024b)	–	74.5	–	86.4	–	51.8	67.7	87.2	82.4	94.0	79.8	85.2
Unified Model												
Blip-3o-4B (Chen et al., 2025b)	73.8	60.4	78.6	–	–	46.6	–	–	78.0	–	–	–
Emu3-8B (Wang et al., 2024)	68.2	57.4	58.5	70.0	–	31.6	47.6	–	64.7	76.3	–	68.7
Janus-Pro-7B (Chen et al., 2025d)	72.1	–	79.2	68.1†	–	41.0	42.5	25.8†	45.6†	40.8†	21.3†	59.0†
X-Omni-7B (Geng et al., 2025)	74.1	62.6†	74.8	76.8†	–	47.2†	54.1†	81.5†	77.4†	88.6	46.9†	70.4†
Bagel-14B (Deng et al., 2025)	78.5†	72.8†	85.0†	89.2†	–	55.3	73.1	78.5†	80.0†	88.1†	51.0†	73.3†
Manzano-3B	74.3	65.1	78.1	82.2	92.9	51.4	69.8	88.2	80.1	93.5	75.0	85.7
Manzano-30B	76.0	70.1	83.4	86.0	96.2	57.8	73.3	89.0	84.4	94.3	81.9	86.3
GPT-4o (Islam & Moushi, 2024)	77.1	75.4	–	84.6	90.7	69.2	61.3	85.7	–	92.8	–	73.6
Gemini-2.5-Pro (Comanici et al., 2025)	–	78	86.3	89.5	88.4	81.7	82.7	83.3	76.8	94.0	84.3	86.2

Table 2: **Comparison on general, knowledge, and text-rich benchmarks**. GPT-4o, Gemini-1.5-Pro, and Gemini-2.5-Pro numbers are from OpenVLM Leaderboard. † represents that the results were reproduced independently in our experiments and might differ from those reported in previous studies. Manzano demonstrate competitive understanding capabilities, especially on text-rich benchmarks.

substantially (+9.9), while instruction following performance remains unchanged. A slight decrease is observed in aesthetic quality. For automatic generation benchmarks, performance on GenEval and DPGEval remains nearly identical, whereas WISE exhibits a modest improvement (+2.0).

Takeaways. Scaling the unified LLM backbone consistently improves both understanding and generation, with substantial gains on text-rich understanding tasks and on WISE for generation. Scaling the image decoder also enhances image quality, without negatively affecting understanding. We observed that performance on the GenEval and DPG benchmarks becomes saturated when the model becomes larger. This saturation motivates a re-examination of how emergent capabilities of unified models could be assessed, as existing benchmarks may capture only a limited portion of overall capability and can be boosted through targeted data tuning (Liu et al., 2025). Meanwhile, we observe substantial improvements on world-knowledge generation tasks, and we hope these findings pave the way for new directions in future community research.

4.4 Comparisons with Unified and Specialist models

To comprehensively assess our Manzano model's capabilities, we compare its performance against SOTA unified and specialist models (i.e., understanding-only and standalone generation models).

4.4.1 Image Understanding

As mentioned in Sec. 4.1, we evaluate our model's understanding capabilities from three perspectives: Knowledge & Reasoning, General Visual Question Answering, and Text-rich Document & Chart Understanding. The results, shown in Table 2, compare our model against other understanding-only models of a similar size. Despite being a unified model, our model achieves state-of-the-art performance on many understanding benchmarks: (i) *Knowledge & Reasoning*: At the 3B scale, our model outperforms all unified models within the 7B scale and achieves performance on par with or better than the best specialist models at the 3B size. At the 30B scale, our model ranks first on the ScienceQA, MMMU, and MathVista benchmarks and third on the AI2D benchmark, outperforming all other unified and specialist models in these categories. Notably, our model surpasses the proprietary models listed in the final three rows on ScienceQA and is competitive with the current state-of-the-art model on the AI2D benchmark. (ii) *General Visual Question Answering*: For general visual question answering, our model generally outperforms other unified models, despite its smaller size. It also achieves competitive results with state-of-the-art specialist models at both scales. (iii) *Text-rich Document and Chart Understanding*: On text-rich and chart understanding tasks, our model achieves the best performance on four out of five benchmarks (ChartQA, TextVQA, DocVQA, and OCRBench) when compared to all other unified, specialist, and proprietary models. For the InfoVQA

Model	GenEval Benchmark							WISE Benchmark						
	Single	Two	Counting	Colors	Position	Color Attr.	Overall	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Dedicated T2I Model														
SDXL-3.5B (Podell et al., 2023)	0.98	0.74	0.39	0.85	0.15	0.23	0.55	0.43	0.48	0.47	0.44	0.45	0.27	0.43
DALL-E 3 (OpenAI, 2024)	0.96	0.87	0.47	0.83	0.43	0.45	0.67	–	–	–	–	–	–	–
SD3-Medium-2B (Esser et al., 2024)	0.99	0.94	0.72	0.89	0.33	0.60	0.74	0.43	0.50	0.52	0.41	0.53	0.33	0.45
PixArt-Alpha-0.6B (Chen et al., 2023)	–	–	–	–	–	–	–	0.45	0.50	0.48	0.49	0.56	0.34	0.47
FLUX.1-dev-12B (Labs, 2024)	0.98	0.93	0.75	0.93	0.68	0.65	0.82	0.48	0.58	0.62	0.42	0.51	0.35	0.50
LLM & Diffusion Conjunction														
MetaQuery-XL-7B (Pan et al., 2025)	–	–	–	–	–	–	0.80†	0.56	0.55	0.62	0.49	0.63	0.41	0.55
OmniGen2-7B (Wu et al., 2025c)	1.00	0.95	0.64	0.88	0.55	0.76	0.80	–	–	–	–	–	–	–
Qwen-Image-27B (Wu et al., 2025a)	0.99	0.92	0.89	0.88	0.76	0.77	0.87	0.67	0.67	0.80	0.62	0.79	0.41	0.67
Unified Multimodal LLM														
Janus-Pro-7B (Chen et al., 2025d)	0.99	0.89	0.59	0.90	0.79	0.66	0.80†	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Bagel-14B-A7B (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82	0.44	0.55	0.68	0.44	0.60	0.39	0.52
X-Omni-7B (Geng et al., 2025)	0.98	0.95	0.75	0.91	0.71	0.68	0.83†	–	–	–	–	–	–	–
Manzano-3B	0.98	0.91	0.82	0.71	0.78	0.71	0.85	0.42	0.51	0.59	0.45	0.51	0.32	0.46
Manzano-30B	1.00	0.91	0.83	0.87	0.84	0.65	0.85	0.58	0.50	0.65	0.50	0.55	0.32	0.54
GPT-4o (Islam & Moushi, 2024)	0.99	0.92	0.85	0.92	0.75	0.61	0.84	0.81	0.71	0.89	0.83	0.79	0.74	0.80

Table 3: **Comparison on GenEval and WISE benchmarks**. † represents the evaluation that involves LLM rewriting. Manzano achieves competitive performance compared with other unified models.

task, our model significantly outperforms its unified counterparts and achieves the best results among specialist models.

4.4.2 Image Generation

We present the quantitative results for our model's image generation capabilities, evaluating them on two benchmarks: GenEval (Ghosh et al., 2023) and WISE (Niu et al., 2025). While both benchmarks assess how well models follow text instructions, WISE additionally evaluates semantic grounding through world-knowledge-informed attributes. As shown in Table 3, our model achieves SOTA results among unified MLLMs on both GenEval and WISE. The 3B model can already perform competitively with or better than much larger unified models, and scaling to 30B further improves generation quality – most notably yielding a large gain on WISE, while maintaining strong GenEval performance. This confirms that our unified architecture and training recipe support strong instruction-following generation. Furthermore, we show the editing capabilities of our model in Sec. B.

4.4.3 Comparison with Unified Models

In addition to specialist models, we also compare against recent unified models such as Janus-Pro (Chen et al., 2025d), X-Omni (Geng et al., 2025), and Bagel (Deng et al., 2025), which aim to handle both understanding and generation within a single framework. Our Manzano model substantially outperforms these unified baselines across almost all understanding benchmarks. At a similar scale, our 3B model exceeds X-Omni and BAGEL on DocVQA, OCRBench, and SEEDBench while maintaining competitive performance on MathVista and ChartQA. Our 30B model further extends this lead, consistently surpassing all existing unified models across knowledge, general VQA, and text-rich domains. This demonstrates that unification does not have to come at the cost of understanding capability. With careful architectural and training design, our model matches or surpasses the best specialist models while providing strong generative capability. We provide more qualitative comparison to the state-of-the-art unified models in Fig. 8.

5 Conclusion

We introduced Manzano, an MLLM that combines visual understanding and image generation through a hybrid image tokenizer and a unified autoregressive backbone. The LLM predicts high-level semantics in the form of text and image tokens, while a lightweight diffusion-based image decoder renders final pixels from the generated image tokens. Coupled with a streamlined three-stage training recipe, this architecture delivers: (i) state-of-the-art on understanding tasks, (ii) substantial gains on generation among unified models, and (iii) minimal task interference as validated by interplay and scaling ablations. Beyond generation, Manzano naturally supports image editing by conditioning both the LLM and image decoder on a reference image, enabling instruction-following with pixel-level control.

REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

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

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Chen Chen, Rui Qian, Wenze Hu, Tsu-Jui Fu, Jialing Tong, Xinze Wang, Lezhi Li, Bowen Zhang, Alex Schwing, Wei Liu, et al. Dit-air: Revisiting the efficiency of diffusion model architecture design in text to image generation. *arXiv preprint arXiv:2503.10618*, 2025a.

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. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025b. URL `https://arxiv.org/abs/2505.09568`.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Tianyi Zhou, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3o-next: A next-generation multimodal foundation model, Aug 2025c. URL `https://jiuhaichen.github.io/BLIP3o-NEXT.github.io/`.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024a.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024c.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2024. URL https://sharegpt4o.github.io/.

Google DeepMind. Create and edit images with gemini. Google DeepMind, 2025. https://deepmind.google/models/gemini/image/.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ben Egan, Alex Redden, XWAVE, and SilentAntagonist. Dalle3 1 Million+ High Quality Captions, May 2024. URL https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions.

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

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

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

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

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

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025a.

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

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

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

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pp. 111–127. Springer, 2024.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *NeurIPS*, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *ICLR*, 2021.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023b.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

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

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

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024c.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

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

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

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

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

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025.

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

OpenAI. Dall·e 3. `https://openai.com/index/dall-e-3/`, 2024.

OpenAI. Addendum to gpt-4o system card: 4o image generation, 2025. Accessed: April 2, 2025.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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

Alec Radford, 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, 2021a. URL `https://arxiv.org/abs/2103.00020`.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Team Seed. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678, 2023.

Qwen team. Qwen2-vl, August 2024. URL https://qwenlm.github.io/blog/qwen2-vl/.

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

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024a.

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

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

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

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Chris Knutsen, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *ICLR*, 2025.

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, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

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

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

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025d.

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

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.

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*, 2022.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

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

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024b.

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

Hanzhi Zhou, Erik Hornberger, Pengsheng Guo, Xiyou Zhou, Saiwen Wang, Xin Wang, Yifei He, Xuankai Chang, Rene Rauch, Louis D'hauwe, et al. Apple intelligence foundation language models: Tech report 2025. *arXiv preprint arXiv:2507.13575*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

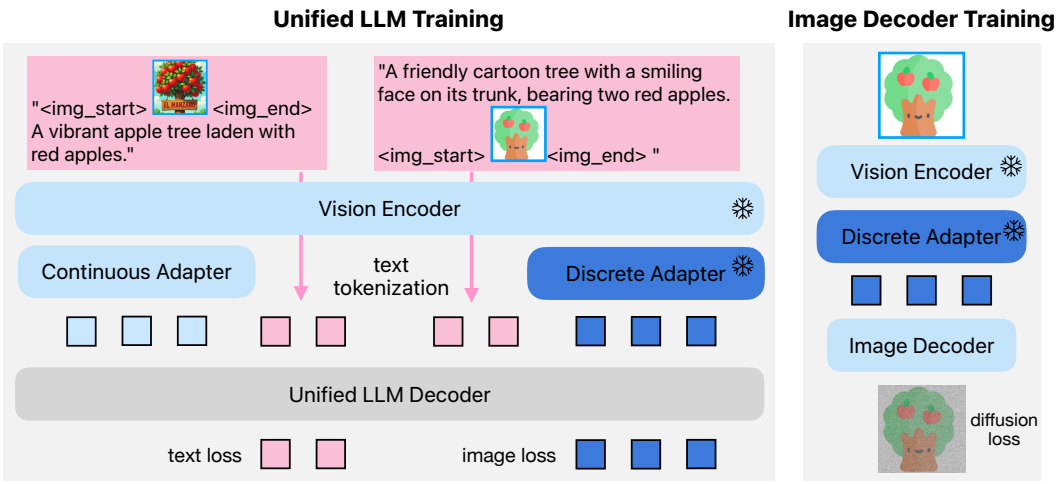


Figure 5: **Training overview**. **(Left)**: Unified LLM training with hybrid tokens, the continuous adapter produces embeddings used for the text loss, while the discrete adapter generates hard tokens serving as targets for the image loss. **(Right)**: With vision encoder and adapters fixed, an image decoder is trained to reconstruct images using a diffusion loss.

A Training

A.1 Data

Our training data mixture includes text-only, image understanding, and generation data, divided into pre-training, continued pre-training, and supervised fine-tuning (SFT) stages. We leverage high-quality text-only data (Zhou et al., 2025) for both pre-training and SFT to maintain the language modeling capability of Manzano model.

A.1.1 Pre-training & Continued Pre-training

Understanding. We use two types of image understanding data: captioning (paired images and text descriptions), and interleaved image-text data. For captioning, we use a combination of sources with 2.3B image-text pairs, including CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), COYO (Byeon et al., 2022), VeCap (Lai et al., 2024), and in-house licensed data. This data undergoes a filtering and re-captioning process to ensure high quality. For interleaved data, we use 1.7B documents from (Laurençon et al., 2024) and web-crawled interleaved data, similar to MM1 (McKinzie et al., 2024) and MM1.5 (Zhang et al., 2024a).

In the continued pre-training stage, we further train on 24M high-quality capability-oriented data, including documentation, charts, multilingual OCR, knowledge & reasoning, high-quality synthetic captions data, all with image splitting (Lin et al., 2023; Gao et al., 2024; Zhang et al., 2024a) enabled.

Generation. The image generation pre-training data consists of 1B in-house text-to-image pairs. Following (Chen et al., 2025a), we generate synthetic captions using different captioner models. For the continued pre-training stage, we select a high-quality subset of licensed images and re-caption them with a more powerful MLLM, generating descriptions of lengths varying from 20 to 128 tokens.

A.1.2 Supervised Fine-tuning

Understanding. Following MM1.5 (Zhang et al., 2024a), our final understanding SFT recipe comprises 75% image–text data and 25% text-only data. The image–text portion is further composed of approximately 30% general knowledge data, 20% document and chart understanding data, and 25% vision chain-of-thought (CoT) and in-house generated reasoning data.

Generation. Our text-to-image SFT data includes a curated blend of real and synthetic data. We begin with real-world text-image pairs from the DreamO dataset (Mou et al., 2025). However, we

observed that training solely on this dataset, while sufficient for standard diffusion-based generators, caused our unified auto-regressive model to overfit. To mitigate this, we expand our training data with synthetic examples. First, we incorporated 90K text-image pairs from established datasets, including DALLE3-1M (Egan et al., 2024), BLIP-3o (Chen et al., 2025b), and ShareGPT-4o (Cui et al., 2024). Second, to achieve a larger scale, we generated an additional 4M pairs by feeding prompts from JourneyDB (Sun et al., 2023) into an open-source standalone diffusion model, Flux.1-schnell (Labs, 2024).

A.2 Training Recipes

A.2.1 Hybrid tokenizer training

The hybrid image tokenizer aims to produce two types of tokens: *continuous* for understanding and *discrete* for generation, which are pre-aligned with the multimodal LLM semantic space.

We first pre-train the vision encoder (ViT) using CLIP Radford et al. (2021b). Then we attach a pretrained small LLM decoder (300M) to the shared vision encoder through two parallel continuous and discrete adapters (See Fig. 2-Left). For each training sample, we randomly select one adapter and feed the corresponding embeddings to the LLM decoder, which is trained with next-token prediction. We unfreeze all parameters and train the model on a variety of understanding data domains, including general knowledge, reasoning, and text-rich tasks.

This process enhances the tokenizer's understanding capability, encompassing both high-level semantic understanding and fine-grained spatial details. Meanwhile, the branches are also being aligned to the same space. We follow the pre-training, continued pre-training and SFT stages using the understanding and text-only data described in Sec. A.1.

After training, we discard the small LLM decoder and retain the resulting hybrid image tokenizer, which is then used as a vision input module for the unified LLM and image decoder.

A.2.2 Unified LLM Training

As shown in Fig. 5-Left, we freeze the parameters of both the vision encoder and the discrete adapter to maintain a fixed vocabulary of image tokens during training. We extend the LLM embedding table with 64K Image tokens following the same codebook size of FSQ layer in the tokenizer.

For image understanding, the image tokenizer extracts the continuous features from the input image and feeds them directly into the LLM with standard next-token loss on text targets. For image generation, the tokenizer uses its discrete adapter to convert input images into a sequence of discrete image token IDs that are mapped to image tokens via the extended LLM embedding table. The LLM then computes a cross-entropy loss on these image tokens only. To balance the training of understanding and generation tasks, we set the weight ratio of text loss to image loss at 1:0.5.

We train the unified LLM in three stages. Pre-training and continued pre-training use a 40/40/20 mix of image understanding, image generation and text-only data as described in Sec. A.1.1. We train our model with 1.6T tokens (0.8T tokens for the 30B model) during the pre-training and an additional 83B tokens during the continued pre-training. Similarly, SFT stage uses curated instruction data with a 41/45/14 mix ratio for understanding, generation, and text using datasets in Sec. A.1.2.

A.2.3 Image Decoder Training

Our image decoder is trained following a progressive resolution growing paradigm (Esser et al., 2024; Chen et al., 2025a). We first pre-train the decoder at a resolution of 256x256 for 400K steps. Subsequently, the model is fine-tuned progressively on higher resolutions of 512, 1024, and 2048, with each stage trained for a shorter schedule of 100K steps. For each stage, only images with short sides larger than the target resolution were used for training.

B Capability Extension to Image Editing

Image editing represents both a crucial application and a natural extension of text-to-image generation. Despite Manzano demonstrating strong multimodal modeling capabilities, particularly on text-rich

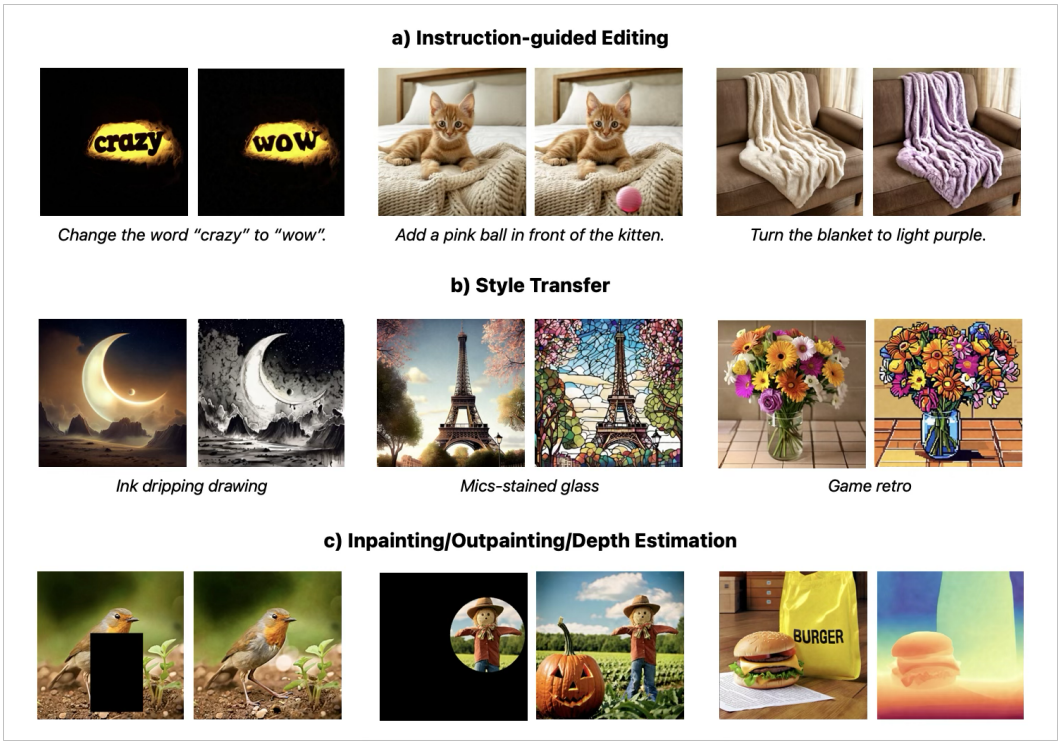


Figure 6: **Editing capabilities of Manzano**. (a) instruction-guided editing, (b) style transfer across diverse visual domains, and (c) extended editing tasks including inpainting, outpainting, and depth-estimation. Manzano achieves pixel-level controls across these five editing tasks.

understanding benchmarks, achieving pixel-level precision in fine-grained image editing remains challenging. Similarly, recent work within the decoupled LLM–diffusion paradigm (Wu et al., 2025c) reports difficulties when relying solely on the LLM for precise editing, since the LLM lacks native mechanisms for direct pixel-level control.

We adopt an approach similar to (Wu et al., 2025c) by providing the reference image simultaneously to both the LLM and the diffusion decoder. In this formulation, the LLM is responsible for diverse instruction following and maintaining semantic coherence, while the diffusion decoder enforces precise pixel-level control. By jointly conditioning on the reference image, Manzano enables accurate semantic instruction following while preserving fine-grained visual consistency. In Fig. 6, Manzano demonstrates versatile editing capabilities, including instruction-guided editing, style transfer, inpainting, outpainting, and depth estimation.

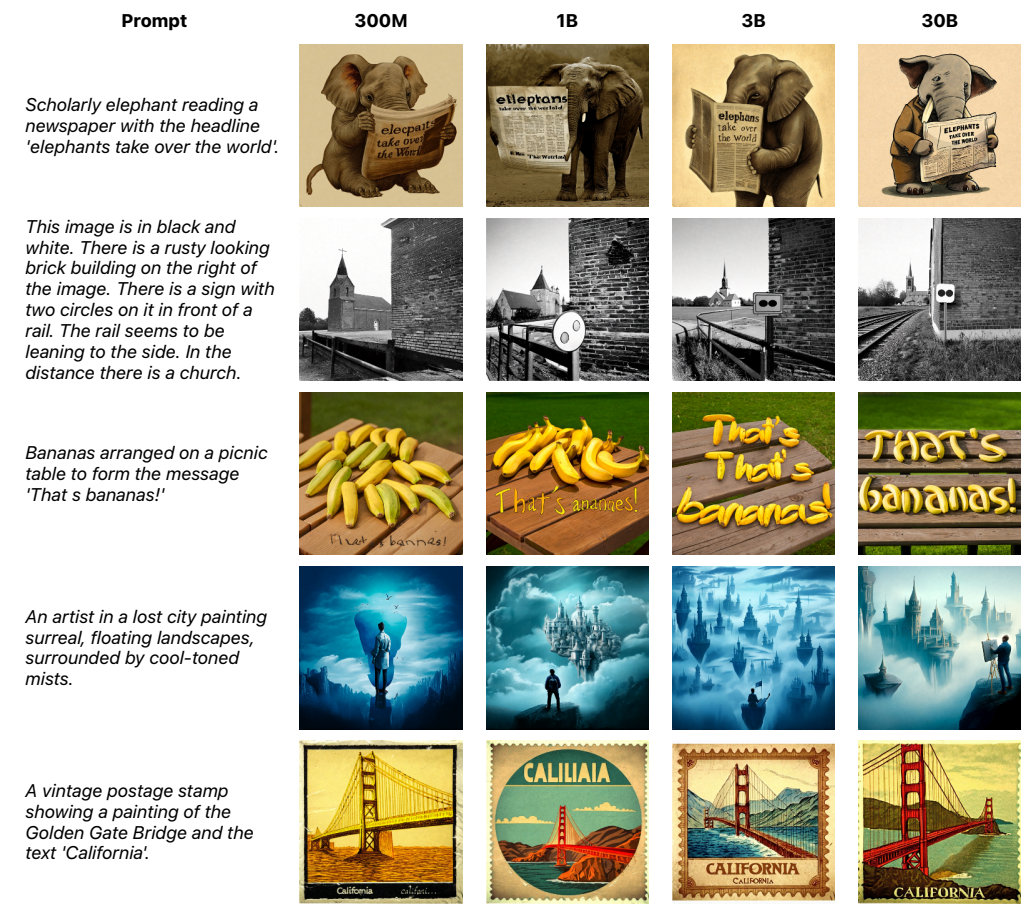


Figure 7: **Qualitative generation results when scaling LLM decoder size**. The generated image quality improves as the LLM decoder size increases. For example, in rows 1, 3, and 5, there is a clear trend toward better text rendering and creativity. In row 2, the scene configuration improves significantly with each increase in the LLM decoder's scale. The 300M model generates an image with only the brick building and the church that are mentioned in the prompt, but as the model grows to 1B and 3B, it begins to include the sign with two circles. Furthermore, the 30B model generates an image that accurately depicts and integrates all the concepts mentioned in the prompt.

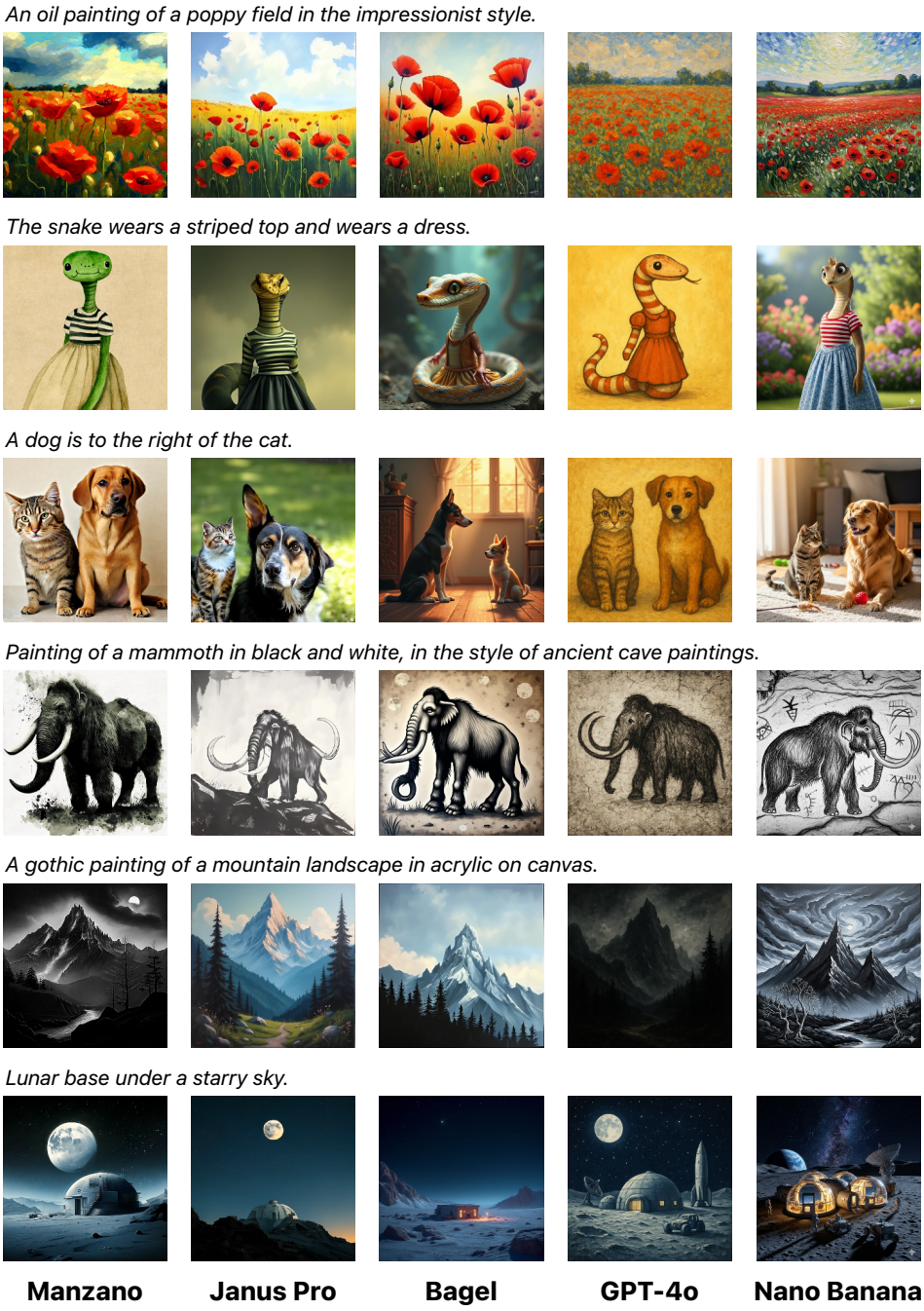


Figure 8: **Qualitative comparison with SOTA unified models**. We compare our Manzano-30B model to the SOTA models through side-by-side comparison. The images generated by our model demonstrate strong capabilities in instruction following, aesthetics, and creativity, often with a photo-realistic quality.