OneCAT: Decoder-Only Auto-Regressive Model for Unified Understanding and Generation

Han Li[1,2*], Xinyu Peng[1,2*], Yaoming Wang[1¶†], Zelin Peng[1,2‡], Xin Chen[1]
Rongxiang Weng[1], Jingang Wang[1¶], Xunliang Cai[1], Wenrui Dai[2¶], Hongkai Xiong[2]

[1]Meituan Inc, [2]Shanghai Jiao Tong University,

*Equal contribution, ¶Corresponding Author, †Project lead

Abstract

We introduce OneCAT, a unified multimodal model that seamlessly integrates understanding, generation, and editing within a novel, pure decoder-only transformer architecture. Our framework uniquely eliminates the need for external components such as Vision Transformers (ViT) or vision tokenizer during inference, leading to significant efficiency gains, especially for high-resolution inputs. This is achieved through a modality-specific Mixture-of-Experts (MoE) structure trained with a single autoregressive (AR) objective, which also natively supports dynamic resolutions. Furthermore, we pioneer a multi-scale visual autoregressive mechanism within the Large Language Model (LLM) that drastically reduces decoding steps compared to diffusion-based methods while maintaining state-of-the-art performance. Our findings demonstrate the powerful potential of pure autoregressive modeling as a sufficient and elegant foundation for unified multimodal intelligence. As a result, OneCAT sets a new performance standard, outperforming existing open-source unified multimodal models across benchmarks for multimodal generation, editing, and understanding.

Date: September 4, 2025
Corresponding: wangyaoming03@meituan.com;wangjingang02@meituan.com;daiwenrui@sjtu.edu.cn
Project Page: https://onecat-ai.github.io/

1 Introduction

In the past few years, modular approaches, utilizing separate architectures for understanding [2, 3, 12–14, 65], generation [20, 24, 34, 36, 40, 59], and editing tasks [5, 32, 46, 85–87], dominated multi-modal frameworks. While producing capable systems, it inherently creates complex, multi-stage pipelines. Such designs often suffer from architectural bottlenecks that limit deep, early-stage fusion of cross-modal information and introduce significant inference latency, presenting a major barrier to both efficiency and performance. In response to these limitations, the field has seen a rapid convergence towards unified multimodal modeling, aiming to integrate these disparate capabilities within a single, end-to-end architecture [11, 16, 62, 67, 75, 76]. Despite the trend towards unification, many current models remain tethered to the modular paradigm. We contend that a true paradigm shift—necessary to unlock the full potential of unified systems—demands a move towards a more fundamental, first-principles architecture that eschews external components. We therefore propose that a pure decoder-only transformer, trained under a unified objective, provides not just a sufficient, but a more elegant and potent foundation for the next generation of general-purpose multimodal intelligence.

‡Work was done during their internship.

Figure 1 Showcase of the text-to-image generation abilities of the **OneCAT** model.

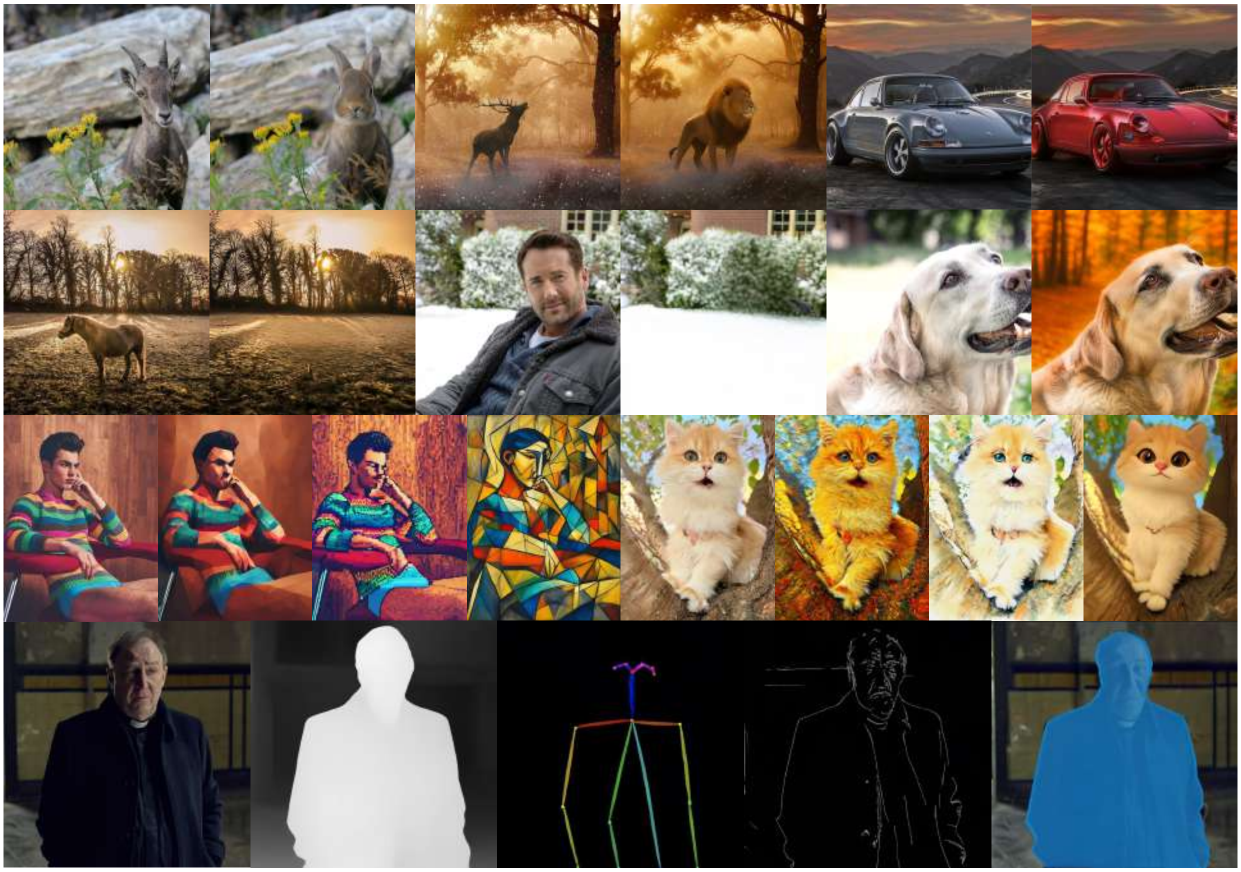

Figure 2 Showcase of the image editing abilities of the **OneCAT** model, including general image editing tasks such as object removal, background adjustment, color adjustment, subject replacement, and style transfer; as well as perceptual tasks including depth estimation, pose estimation, object segmentation, and Canny edge detection.

To realize this vision, we re-evaluate the core architectural tenets of multimodal systems. We introduce an encoder-free framework where raw visual inputs are directly tokenized into patch embeddings and processed alongside text tokens within a single decoder stack. The critical innovation is a modality-specific Mixture-of-Experts (MoE) layer, which dynamically routes vision and text tokens to specialized experts, enabling deep, early-stage feature fusion in an efficient inference manner without the need for exquisite encoders. For generative tasks, we pioneer a multi-scale autoregressive mechanism [24, 63] inside the LLM, augmented with the proposed scale-aware adapter modules, to predict image tokens from low to high resolution. This design not only circumvents the high latency of iterative diffusion models but also allows the model to learn a coarse-to-fine generative process, significantly enhancing both speed and output quality.

Regarding data protocol, we leverage mixed training strategies that combine large-scale, web-scraped image-text pairs with curated, instruction-following datasets. Our pre-training corpus is constructed to expose the model to a diverse range of data formats, including image-text documents, visual question-answering pairs, and generative prompts for both text-to-image synthesis and image editing. This heterogeneous data mixture is crucial for training a truly unified model, as it forces the shared decoder to develop a generalized representation that can seamlessly switch between comprehension, generation, and editing tasks. We argue that this unified data diet, coupled with our bottleneck-free architecture, is key to fostering the synergistic emergence of complex multimodal capabilities.

Building upon these principles, we present Only DeCoder Auto-regressive Transformer (OneCAT), an open-source unified multimodal model. Comprehensive evaluations demonstrate that OneCAT sets a new state-of-the-art for pure decoder-only unified models. More importantly, the encoder-free design provides a significant inference speedup, particularly for high-resolution inputs. By demonstrating the viability and superiority of a pure decoder architecture, OneCAT offers a more first-principles-aligned paradigm for multimodal modeling. It facilitates earlier cross-modal fusion through its unified MoE structure and enhances semantic consistency, providing valuable insights and a powerful new baseline for the development of next-generation unified multimodal systems. For examples of our model's impressive image generation capabilities, please refer to Fig. 1. Its advanced image editing functionalities are further showcased in Fig. 2.

2 Related Work

2.1 Compositional MLLMs

The field of Multimodal Large Language Models (MLLMs) has rapidly evolved, converging on a dominant **compositional architecture**. This paradigm connects a pre-trained vision encoder, such as CLIP [56], SigLIP [84], or the more recent InternViT [14], to a powerful Large Language Model (LLM) through a trainable connector. Pioneering works [1, 37] propose sophisticated connector designs. For example, Flamingo [1] introduces gated cross-attention layers to inject visual information into an LLM, while BLIP-2 [37] develops the Q-Former to bridge the modality gap between an image encoder and an LLM. A significant shift occurs with LLaVA [45], which simplifies the connector to a lightweight Multi-Layer Perceptron (MLP) projection layer. This effective design become a foundational blueprint for numerous subsequent MLLMs. For example, recent state-of-the-art models like the InternVL series [12–14, 66, 91] and the Qwen-VL series [2, 3, 65] adopt this same core principle and achieve superior performance by leveraging larger-scale training data and more powerful vision and language foundation models. However, this successful compositional design has inherent drawbacks. The separate nature of the vision and language components complicates the end-to-end optimization process and introduces two critical bottlenecks. First, the sequential nature of the architecture, where the vision encoder must fully process an image before the LLM can begin its generation, leads to high inference latency. This initial step is often referred to as the "prefilling" stage. Second, the connector itself acts as an information bottleneck. In this so-called **late fusion** pipeline, complex visual information is compressed into a compact representation for the LLM, inevitably causing a loss of fine-grained visual detail. These fundamental limitations are now motivating a shift in the field towards more deeply integrated, such as decoder-only models, that aim to overcome these challenges.

2.2 Decoder-only MLLMs

Decoder-only MLLMs, also known as monolithic MLLMs, have recently emerged as a minimalist yet powerful alternative to the conventional compositional architecture. This paradigm aims to achieve greater efficiency and a more direct **early fusion** of modalities by eliminating the separate vision encoder. For example, Fuyu-8B [4] processes vision patches by feeding them through a simple linear patch embedding layer directly into the LLM backbone, which markedly reduce inference latency. Inspired by this success, subsequent works [18, 19, 49] further advance decode-only MLLMs by targeting their training processes and architectures. EvE [18] introduces a novel training objective that guides the model's visual learning by aligning the last layer of LLM's hidden states of image patch tokens with semantic features from a pre-trained vision encoder. Differently, Mono-InternVL [49] and EvEv2.0 [19] adopt a Mixture-of-Experts (MoE) framework, introducing a dedicated *visual expert* to handle visual-specific features more effectively. HoLVE [61] prepends a causal transformer to the LLM to explicitly convert both visual and textual inputs into a shared space. Despite these promising advancements, the overall training efficiency of these models remains a significant challenge. More importantly, the potential for the decoder-only architecture to create unified models that can seamlessly integrate multimodal understanding, generation, and even image editing capabilities remains a largely unexplored research avenue.

2.3 Unified Visual Understanding and Generation

Building on the success of MLLMs, the convergence of visual understanding and generation into a unified framework now represents a key research frontier. Pioneering unified MLLMs such as Chameleon [62],Transfusion [90], emu3 [67] and show-o [75] utilize visual tokenizers (e.g., VQ-VAE) to convert images into a sequence of discrete tokens, enabling seamless multimodal understanding and generation within a single model. However, the discretization inevitably results in lossy visual information and weakens in extracting semantic contents. Janus series [11, 69] decouples visual encoding for understanding and generation using two separate encoders, but may compromise performance due to task conflicts in shared LLM parameter space. Metaqueries [53], BLIP3-O [8], Uniworld-V [43] assembles off-the-shelf specialized MLMMs and diffusion models by tuning adapters and learnable query tokens, which sacrifices true architectural unification for modularity. BAGEL [16] and Mogao [41] employ a Mixture-of-Transformers (MoT) architecture, dedicating different components to autoregressive text generation and diffusion-based visual generation. However, while powerful, this hybrid approach inherits the significant inference latency of diffusion models and still requires separate encoders and tokenizers during the inference.

In contrast to these approaches, our OneCAT introduces a pure decoder-only architecture. By integrating modality-specific experts directly within the decoder, OneCAT achieves versatile multimodal capabilities without the need for external vision encoders or tokenizers at inference time, thus resolving the trade-off between architectural purity and inference efficiency.

2.4 Next Scale Prediction for Visual Generation

Autoregressive models based on next-token prediction(NTP) have long faced efficiency challenges in high-resolution image generation due to the quadratic growth of sequence length with image size. Similarly, diffusion models—though widely successful—often suffer from slow iterative sampling. To address these limitations, VAR [63] introduced the next-scale prediction(NSP) paradigm, which encodes images into hierarchical discrete tokens via a multi-scale VAE and generates them autoregressively from low to high resolution, significantly reducing the number of decoding steps. Building upon this, Infinity [24] further enhanced this approach with bit-level prediction and extended tokenizer vocabulary, achieving superior generation quality while maintaining efficient inference. To enable unified understanding and generation, VARGPT [92] stack the transformer from pretrained VAR [63] as a visual decoder atop a LLM. However, since the visual tokens (*i.e.,* the input of the visual decoder) must be decoded token-by-token through the LLM before subsequent next-scale prediction, this approach compromises the inference efficiency that is the key advantage of the NSP.

In contrast, our proposed OneCAT seamlessly unifies next-token prediction for text generation and next-scale prediction for visual generation within a single LLM.

3 OneCAT

As illustrated in Fig. 3, OneCAT employs a pure decoder-only architecture, eliminating the need for any additional vision encoder or tokenizer during inference. This streamlined design significantly simplifies the model structure and reduces computational overhead. Unlike traditional multimodal models that rely on semantic encoders like ViTs for understanding [3, 11, 16, 41], OneCAT utilizes a lightweight `Patch Embedding` layer. This layer losslessly converts raw images into continuous visual tokens, enabling efficient multimodal understanding. Crucially, the same `Patch Embedding` layer also encodes reference images for editing tasks, thereby superseding separate VAE-based tokenizers traditionally used [16, 41] and further enhancing inference efficiency.

At its core, OneCAT integrates a Mixture-of-Experts (MoE) architecture. This MoE comprises three specialized feed-forward network (FFN) experts: one dedicated to processing text tokens for language comprehension, another designed for continuous visual tokens to facilitate multimodal understanding, and a third optimized for generating discrete visual tokens during image synthesis. All QKV and attention layers are **shared across modalities and tasks**, which promotes significant parameter efficiency and robust cross-modal alignment for instruction-following.

For text generation, OneCAT adheres to the conventional **Next-Token Prediction** paradigm, leveraging the well-established capabilities of autoregressive language modeling. In parallel, for visual generation, it innovatively employs the **Next-Scale Prediction** paradigm [24, 63]. This mechanism generates images in a coarse-to-fine, hierarchical manner, progressively predicting visual tokens from the lowest to the highest resolution scale, thereby achieving high-quality visual outputs.

3.1 Architecture

Our OneCAT model is initialized from the pre-trained Qwen2.5 LLM [77], leveraging its strong language modeling capabilities as a foundation. To construct our MoE architecture, we replicate the original FFN layer from each Qwen2.5 transformer block to form three distinct experts: a Text FFN (*i.e.,* `Text. FFN`), a Visual Understanding FFN (*i.e.,* `Und. FFN`), and a Visual Generation FFN (*i.e.,* `Gen. FFN`). OneCAT employs a

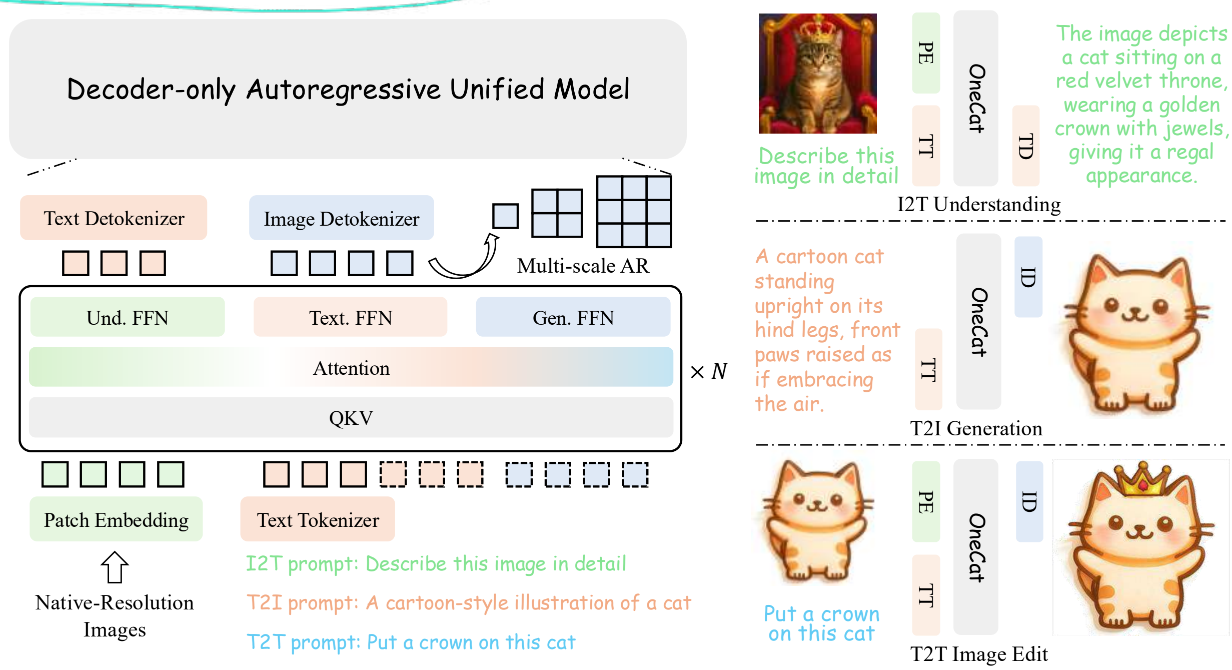

Figure 3 Overview inference pipeline of OneCAT. OneCAT is a decoder-only autoregressive unified model that seamlessly supports multimodal understanding, text-to-image generation and image editing. OneCAT implements both **next-token prediction** for text generation and **next-scale prediction** for visual generation within a unified LLM backbone. In the left part, the dash squares denote the generated text and visual token during inference. In the right part, **PE** and **TT** denote patch embedding and text tokenizer, respectively. **TD** and **ID** denote text detokenizer and image detokenizer, respectively.

deterministic hard routing mechanism, where tokens are assigned to a specific expert based on their modality and the task at hand. The core functionality for each task is handled as follows:

Multimodal Understanding. To process images for understanding tasks, we employ a simple yet effective patch embedding layer that converts raw images into a sequence of continuous visual tokens. This layer consists of a 14×14 convolution for image patchifying, a 2×2 pixel unshuffle operation for visual token compression, and a two-layer Multilayer Perceptron (MLP) for projecting the visual features to match the LLM's hidden state dimension. These continuous visual tokens are exclusively routed to the `Und. FFN`, while the text tokens for instruction are routed to the `Text FFN`.

Text-to-Image Generation. For image generation, we leverage a pre-trained multi-scale VAE model from Infinity [24] to map images between pixel space and latent space. This VAE operates with a downsampling ratio of 16 and a latent channel size of 32, and incorporates a bitwise quantizer [89] to enlarge the vocabulary. During training, the image tokenizer transforms the target images into a sequence of multi-scale discrete visual tokens to serve as ground-truth and input of LLM for teacher-forcing training, which are processed by the `Gen. FFN`. Critically, during inference, the tokenizer is not required; only the detokenizer is needed to reconstruct the final image from the generated multi-scale visual tokens. The conditional text tokens are also routed to the `Text FFN`.

Image Editing. OneCAT seamlessly supports image editing task by conditioning the visual generation process on a reference image and edit-instruction. The reference image is processed through the patch embedding layer, and the resulting continuous visual tokens are routed to the `Und. FFN` to serve as the visual condition, while the text tokens for instruction are routed to the `Text FFN`. The patch embedding layer provides a **near-lossless representation** of the reference image, which allows the LLM's shallower layers to access low-level features for pixel-level consistency, while the deeper layers extract high-level features for semantical comprehension. Guided by this rich, hierarchical visual context, the model then autoregressively predicts new discrete visual

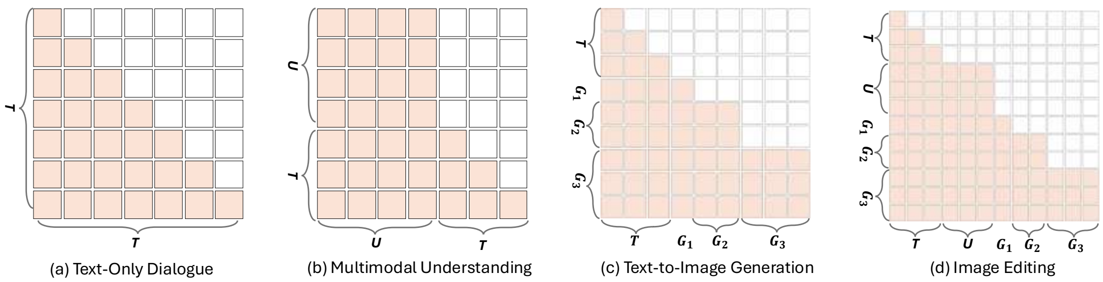

Figure 4 Multimodal versatile attention mechanism. T denotes the text tokens. U denotes the continuous visual tokens for multimodal understanding or reference image tokens for image editing. G_i denotes the i-th scale discrete visual tokens for visual generation.

tokens, which are handled by the `Gen. FFN`. This design enables powerful conditional generation without requiring any architectural modifications, showcasing the versatility of our unified decoder-only design.

3.2 Scale-Aware Adapter for Hierarchical Generation

While the `Gen. FFN` is capable of processing discrete visual tokens, the tokens produced by the multi-scale VAE are inherently hierarchical. A standard FFN would treat tokens for different scales equally, ignoring this crucial structural information. To address this and enable more granular control over the visual generation process, we introduce the **Scale-Aware Adapter (SAA)**, a novel architectural component integrated with the `Gen. FFN`.

Our design is motivated by the principle that different scales in the multi-scale VAE govern distinct aspects for image generation. Specifically, tokens from lower scales predominantly encode low-frequency global information, such as color, illumination, and coarse structure. Those from higher scales, conversely, capture high-frequency details including fine textures and intricate patterns. Processing these functionally divergent tokens with a shared `Gen. FFN` layer limits representational capacity and is thus suboptimal.

The Scale-Aware Adapter (SAA) comprises a set of parallel modules that serve as skip connections over the `Gen. FFN`. Each module is dedicated to processing tokens from a specific scale of the multi-scale VAE, with the total number of adapters matching the number of VAE scales. During inference, discrete visual tokens are routed to their corresponding scale-specific adapter based on the scale index. To ensure parameter efficiency, each adapter is constructed using a low-rank decomposition (rank $r = 64$), inspired by the LoRA [26]. However, unlike LoRA which is typically used for fine-tuning, the SAA modules are trained jointly end-to-end as permanent components of the LLM.

3.3 Multimodal versatile attention mechanism

We leverage a multimodal versatile attention mechanism based on PyTorch FlexAttention [55] to empower OneCAT with the ability to process diverse modalities and tasks in a flexible and adaptive manner. As illustrated in Fig. 4, text tokens T are processed using causal attention, ensuring autoregressive generation; Continuous visual tokens U are processed via full attention, allowing each token to attend to all others in the sequence. Multiscale discrete visual tokens G_i (where i denotes the scale index) are processed via block causal attention, tokens within the same scale can attend to each other freely, while attention across scales follows a causal attention.

4 Model Training Pipeline

The training pipeline is divided into three stages and the instruction of the training recipe is shown in Tab. 1.

4.1 Stage-1: Multilmodal Pretraining

The objective of this stage is to equip the OneCAT with foundational visual perception and generation abilities while maintaining the linguistic capabilities from the pretrained LLM. The core challenge is that, for visual perception, the `Und. FFN` is initialized from the weights of the LLM's text-focused FFN. While this "warm

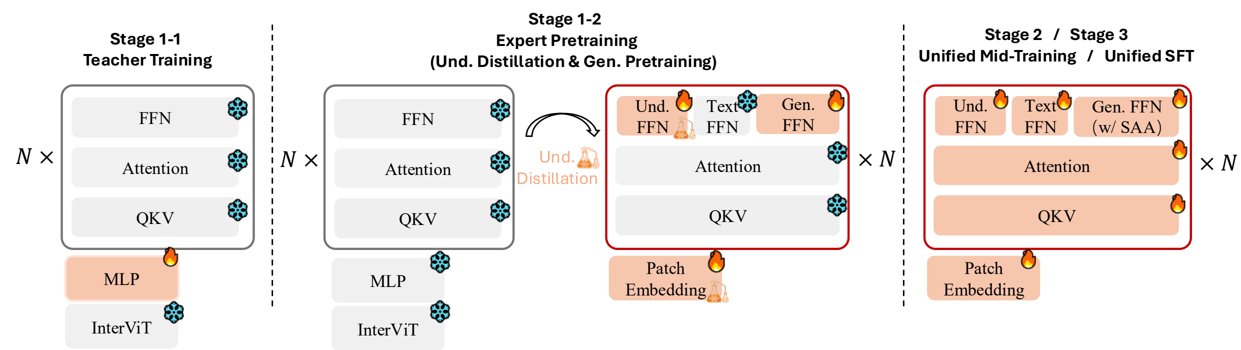

Figure 5 Overview of the training pipeline. In Stage 1, we first prepare a teacher model by training a two-layer MLP to connect InterViT [14] and the Qwen2.5 LLM [77]. This teacher model is then used to perform understanding distillation for the `Und. FFN` and the `Patch Embedding` layer. Simultaneously, we perform generation pretraining to optimize the `Gen. FFN`. All other parameters of the LLM remain frozen to preserve its pretrained language capabilities. In Stage 2 and 3, the entire model is unfrozen to conduct unified mid-training and supervised fine-tuning (SFT), respectively. The VAE component for visual generation is omitted from the figure for clarity.

start" benefits abstract reasoning, it inherently lacks pretrained visual knowledge, making the training process highly data-intensive. To address this limitation, we leverage the visual perception capabilities of an MLLM teacher and introduce an understanding distillation strategy to optimize `Und. FFN` that significantly enhances visual learning efficiency. In parallel, we conduct generation pretraining to optimize `Gen. FFN`.

4.1.1 Stage 1-1: Teacher Training

Rather than employing an off-the-shelf MLLM as teacher (*e.g.,* Qwen2.5-VL [3]), we construct a custom teacher model to ensure parameter consistency between the LLM backbones of the teacher and student models, thereby improving distillation efficiency. Specifically, the teacher is built by connecting a pre-trained vision encoder (`InterViT` [14]) and a LLM (`Qwen2.5` [77]) with a two-layer `MLP` as connector. During this sub-stage, we freeze the ViT and LLM, and exclusively train the MLP connector on a curated small-scale dataset of image-to-text caption pairs using the standard NTP loss.

4.1.2 Stage 1-2: Expert Pretraining

With the teacher model prepared, we proceed to train the OneCAT model. We keep the `QKV Projection`, `Attention`, and `Text FFN` frozen, and selectively optimize the task-specific modules: the `Und. FFN` and `Patch Embedding Layer` for multimodal understanding, and the `Gen. FFN` for text-to-image generation.

Understanding Distillation: We optimize the `Und. FFN` on a large-scale dataset of image-to-text caption pairs to acquire fundamental visual knowledge. Based on the specifically designed teacher, the training objective is a combination of the NTP loss ($\mathcal{L}_{\text{NTP}}$) and a distillation loss ($\mathcal{L}_{\text{Distill}}$). Specifically, $\mathcal{L}_{\text{NTP}}$ is the standard cross-entropy loss for autoregressive text generation. For distillation, instead of matching the final output logits, we align the student's internal hidden states with those of the teacher model through deep feature-level matching. This strategy enables the student to not only mimic the teacher's final prediction but also its intermediate computational patterns across all token positions (including both visual and text tokens) for better visual knowledge transfer. The distillation loss $\mathcal{L}_{\text{Distill}}$ is formulated as the sum of MSE losses between the hidden states of the student and teacher models over all N transformer layers:

$$\mathcal{L}_{\text{Distill}} = \sum_{n=1}^{N} \text{MSE}(\mathbf{h}_S^{(n)}, \mathbf{h}_T^{(n)}), \tag{1}$$

where $\mathbf{h}_S^{(n)}$ and $\mathbf{h}_T^{(n)}$ represent the hidden state outputs from the n-th transformer block of the student and teacher models, respectively. The final objective is thus:

$$\mathcal{L}_{\text{Und}} = \mathcal{L}_{\text{NTP}} + \lambda \mathcal{L}_{\text{Distill}}, \tag{2}$$

Hyperparameter / Config	Stage 1-1 (Teacher Training)	Stage 1-2 (Expert Pretraining)	Stage 2 (Unified Mid-Training)	Stage 3 (Unified SFT)
Learning Rate	2×10^{-3}	2×10^{-4}	2×10^{-5}	1×10^{-5}
LR Scheduler	Cosine	Cosine	Cosine	Cosine
Weight Decay	0	0	0.01	0.01
Gradient Norm Clip	1.0	1.0	1.0	1.0
Batch Size	512	2048	512	256
Sequence Length	1024	1024	8192	16384
Number of Sample: Text-Only	-	-	40M	2M
Number of Sample: Und.	10M	436M	70M	11M
Number of Sample: Gen.	-	52M	60M	3M
Number of Token (Total)	5B	0.3T	0.6T	57B
Token Ratio (T:U:G):	0:1:0	0:8:1	1:2:6	1:5:6
Resolution: Und.	448×448	448×448	Native	Native
Use thumbnail	×	×	✓	✓
Resolution: Gen.	-	256×256	Dynamical (#sides: 288∼864)	Dynamical (#sides: 288∼1728)
Number of Scales : Gen.	-	7	10	10∼13

Table 1 Detailed hyperparameter and configuration of the training recipe across different stages.

where $\lambda = 0.2$ is a balancing hyperparameter. Throughout distillation, all models process images at a fixed resolution of 448×448 to balance computational load and the granularity of visual features.

Generation Pretraining: In parallel, we optimize the `Gen. FFN` on a delicate text-to-image generation dataset to enable the model to learn the spatial relationships and dependencies between multi-scale discrete visual tokens. We adopt the cross-entropy loss for next-scale prediction [24, 63] and the output image resolution is fixed to 256×256 in this stage.

4.2 Stage-2: Unified Mid-Training

In the second stage, we unfreeze the entire model to achieve unified mid-training across multiple tasks (*i.e.*, image-to-text, text-to-image, image editing, and text-only dialogues). At this stage, we incorporate the proposed `scale-aware adapter (SAA)`, which is optimized with other modules of LLM together to extract scale-specific representation for enhanced image generation.

We introduce native resolution strategy for both understanding and generation in this stage. For multimodal understanding, the model is trained to process images at their original resolutions, thereby preserving fine-grained details and eliminating information loss. Additionally, a thumbnail of resolution 448×448 is included to provide global visual context. For visual generation, the model is trained to generate images with different resolution and aspect ratios where the side lengths dynamically sampled from a range of 288 to 864 pixels, which significantly enhancing its generative versatility and real-world applicability.

4.3 Stage-3: Unified Supervised Fine-tuning

The final stage involves unified supervised fine-tuning (SFT) using a curated dateset of higher-quality data to enhance instruction-following and visual generation quality. The native resolution strategy was continued, with the size of generated image expanded to support side lengths between 288 to 1728 pixels, enabling higher-resolution results.

5 Data Setup

Stage-1: For the multimodal understanding, we curate a large-scale dataset of approximately 436 million image-text pairs, which is meticulously compiled and processed through comprehensive filtering and deduplication. This dataset is collected from two primary sources: (1) Public Available Image-Text Caption Pairs: We incorporate several publicly available, high-quality image-caption datasets, including Recap-DataComp-1B [38], Capsfusion [81], Detailed-Caption [39], SA1B-Dense-Caption [15], and Moondream2-COYO-5M-Captions [31].

	OneCAT-1.5B	OneCAT-3B
Base Model	Qwen2.5-1.5B-instruct	Qwen2.5-3B-instruct
Active Parameters	1.5B	3B
Total Parameters	4.5B	9B
Understanding Distillation		
Teacher ViT	InterViT-300M	
Teacher LLM	Qwen2.5-1.5B-instruct	Qwen2.5-3B-instruct

Table 2 Model configurations for the two variants of OneCAT.

(2) Re-captioned Image Datasets: We generate new image-text pairs by re-captioning large-scale public image collections. The source image datasets for this process include COYO700M [6], CC12M [7], CC3M [58], LAION-400M [57], and Zeor250M [74]. From this dataset, we randomly select a subset of 10 million samples to train the custom teacher.

For image generation, we construct a dataset of 52 million text-to-image samples after a rigorous filtering process to remove samples with low resolution or poor aesthetic scores. This collection consists of 1 million class-labeled images from ImageNet-1k [17], 20 million pairs from re-captioned public collections (*i.e.,* COYO700M [6], LAION-400M [57] and CC12M [7]), and 30 million in-house synthetic images generated by FLUX. The overall training token ratio across multimodal understanding and visual generation samples in Stage-I is approximately **8:1**.

Stage-2: In the unified mid-training, for multimodal understanding we leverage an in-house dataset of 70 million visual instruction samples. This dataset is specifically curated to be highly diverse tasks, including general VQA, detailed image captioning, OCR, multimodal reasoning(*i.e.,* STEM problem-solving), knowledge, and visual grounding. For visual generation, we supplement the text-to-image samples of Stage-1 with a additional collection of 8 million image editing samples, resulting a total of 60 million visual generation samples. These additional image editing samples are sourced from several public image editing datasets, including AnyEdit [80], UltraEdit [88], HQ-Edit [30] and OmniEdit [68]. Additionally, we incorporate 40 million text-only instruction samples to preserve the language ability of LLM. To ensure a strong focus on visual generation in Stage-II, we oversample the visual generation data, resulting a final training token ratio of approximately **1 :2 :6** across text-only, multimodal understanding, and visual generation tasks, respectively.

Stage-3: In the SFT stage, for multimodal understanding and text-only instruction, we construct a high-quality dataset of 13 million samples, combining 10 million from public MAmmoTH-VL dataset [22] with 3 million of our in-house synthetic instruction samples to further improve the reasoning abilities. For visual generation, we utilize a total of 3 million samples, aggregated from UniWorld [44], blip3o-60k [8], ShareGPT-4o-Image [10], and additional in-house synthetic data generated by GPT-4o and FLUX. The overall training token ratio across text-only, multimodal understanding, and visual generation is approximately **1 :5 :6**.

6 Implementation details

Model Configurations: We conduct our experiments on two model variants, OneCAT-1.5B and OneCAT-3B. The OneCAT-1.5B model is based on Qwen2.5-1.5B-instruct [3] and contains 1.5B active parameters (4.5B total). The OneCAT-3B model is built upon Qwen2.5-3B-instruct [3] and utilizes 3B active parameters (9B total). For understanding distillation, the ViT of teacher model is InterViT-300M [14] for both two variants and the LLM of teacher model is Qwen2.5-1.5B-instruct and Qwen2.5-3B-instruct for OneCAT-1.5B and OneCAT-3B, respectively, to align with the LLM backbone of our OneCAT. Detailed instruction of the model configurations is shown in Tab. 2.

Data Packing and Gradient Accumulation: To optimize workload balance across distributed processes and increase training throughput, we employ a data packing strategy that concatenates multiple variable-length samples into contiguous sequences. Furthermore, to manage the gradient contributions and token ratios between modalities as in Table 1, we utilize an *uneven* gradient accumulation strategy: prior to each optimizer step, we accumulate a *distinct* number of micro-batches' gradients for the text and image generation tasks to

Model	# Params		TextVQA↑	ChartQA↑	InfoVQA↑	DocVQA↑	GQA↑	AI2D↑
	A-LLM	Vis.						
Encoder-based Understanding Only Models								
InternVL2-2B [13]	1.8B	0.3B	73.4	76.2	58.9	86.9	-	74.1
InternVL2.5-2B [12]	1.8B	0.3B	74.3	79.2	60.9	88.7	-	74.9
Qwen2-VL-3B [3]	1.5B	0.6B	79.7	73.5	65.5	90.1	-	74.7
Qwen2.5-VL-3B [3]	3B	0.6B	79.3	84.0	77.1	93.9	-	81.6
Encoder-free Understanding Only Models								
Mono-InternVL [49]	1.8B	/	72.6	73.7	-	-	59.5	68.6
EvE [18]	7B	/	56.8	59.1	-	-	62.6	61.0
EvEv2 [19]	7B	/	71.1	73.9	-	-	62.9	74.8
HoVLE [61]	2.6B	/	70.9	78.6	55.7	86.1	64.9	73.0
Unified Models								
Chameleon [62]	7B	-	4.8	2.9	5.0	1.5	-	46.0
Emu3 [67]	8B	0.3B	64.7	68.6	43.8	76.3	60.3	70.0
Harmon-1.5B [71]	1.5B	0.9B	-	-	-	-	58.9	-
Show-o2-1.5B [76]	1.5B	0.5B	-	-	-	-	60.0	69.0
Janus-Pro-1.5B [11]	1.5B	0.3B	-	-	-	-	59.3	-
OneCAT-1.5B	1.5B	/	<u>67.0</u>	<u>76.2</u>	<u>56.3</u>	<u>87.1</u>	60.9	72.4
ILLUME+ [28]	3B	0.6B	-	69.9	44.1	80.8	-	74.2
VILA-U [72]	7B	0.4B	60.8	-	-	-	60.8	-
Janus-Pro-7B [11]	7B	0.3B	-	-	-	-	62.0	-
Tar-7B [23]	7B	0.4B	-	-	-	-	61.3	-
Show-o2-7B [76]	7B	0.5B	-	-	-	-	<u>63.1</u>	**78.6**
OneCAT-3B	3B	/	**73.9**	**81.2**	**64.8**	**91.2**	**63.1**	<u>77.8</u>

Table 3 Performance comparison across multiple multimodal understanding benchmarks. Higher scores are better, as indicated by the up-arrow (↑). **A-LLM** denotes the number of activated LLM parameters, while **Vis.** indicates the parameter count of the vision encoder or tokenizer for multimodal understanding. Chameleon [62] does not report the parameter count of its vision tokenizer. slash (/) denotes that models do not require a vision encoder or tokenizer for multimodal understanding. Top-1 accuracy is reported (Best in **bold**, second best is <u>underlined</u>).

obtain a gradient of desired token ratios. Such an approach provides fine-grained control over the effective batch sizes of different tasks, ensuring a balanced and stable joint-training.

Unbiased Global Batch Gradients: When training on N distributed processes, naively averaging local loss can lead to biased gradients when per-process token counts vary. The ideal objective is to optimize the *Global Batch Loss*, defined as the loss summed over tokens for all micro-batches, normalized by the global token count, denoted as T_{global}. To this end, we first prefetch all micro-batches for the next optimizer step, enabling each process to compute the local token counts; a subsequent *All-Reduce* collective operation then aggregates these local token counts into the final global token count, *i.e.*, T_{global}. Similar to [42], we then employ *Global Batch Reduced Loss* by dividing each micro-batch loss by the averaged token count per process, $\frac{T_{global}}{N}$, which can be shown that the final synchronized gradient for the subsequent optimizer step is mathematically equivalent to the gradient of the global batch loss, enabling training with unbiased gradients.

7 Evaluation

7.1 Multimodal understanding.

We evaluate our model on 12 public Multimodal understanding benchmarks spanning diverse capabilities: MMbench [47], MME [79], MMMU [83], MM-Vet [82], and SEED [35] assess general multimodal perception and reasoning. MathVista [48] focuses on mathematical reasoning. TextVQA [60], ChartQA [50], InfoVQA [52], and DocVQA [51] evaluate OCR and text-related visual question answering; while GQA [29] evaluates visual scene understanding and AI2D [33] evaluates scientific diagram comprehension.

As shown in Tab. 3 and Tab. 4, we compare OneCAT with three types of models: encoder-based understanding-only models, encoder-free understanding-only models, and unified MLLMs. Our OneCAT-3B model demon-

Model	# Params		MME-P↑	MME-S↑	MMBench↑	MMMU↑	MM-Vet↑	MathVista↑	SEED↑
	A-LLM	Vis.							
Encoder-based Understanding Only Models									
InternVL2 [13]	1.8B	0.3B	1440	1877	73.2	34.3	44.6	46.4	71.6
InternVL2.5 [12]	1.8B	0.3B	-	2138	74.7	43.6	60.8	51.3	-
Qwen2-VL [65]	1.5B	0.6B	-	1872	74.9	41.1	49.5	43.0	-
Qwen2.5-VL [3]	3B	0.6B	-	2157	79.1	53.1	61.8	62.3	-
Encoder-free Understanding Only Models									
Mono-InternVL [49]	1.8B	/	-	1875	65.5	33.7	40.1	45.7	67.4
EvE [18]	7B	/	-	1628	52.3	32.6	25.7	-	64.6
EvEv2.0 [19]	7B	/	-	1709	66.3	39.3	45.0	-	71.4
HoVLE [61]	2.6B	/	-	1864	71.9	33.7	44.3	46.2	70.7
Unified Models									
Chameleon [62]	7B	-	-	-	35.7	28.4	8.3	-	30.6
Emu3 [67]	8B	0.3B	-	-	58.5	31.6	37.2	-	68.2
Harmon [71]	1.5B	0.9B	1155	1476	65.5	38.9	-	-	67.1
Show-o2 [76]	1.5B	0.5B	1450	-	67.4	37.1	-	-	65.6
Janus-Pro [11]	1.5B	0.3B	1444	-	75.5	36.3	39.8	-	-
OneCAT-1.5B	1.5B	/	1509	1893	72.4	39.0	42.4	55.6	70.9
ILLUME+ [28]	3B	0.6B	1414	-	**80.8**	<u>44.3</u>	40.3	-	**73.3**
VILA-U [72]	7B	0.4B	1401	-	-	-	33.5	-	59.0
Janus-Pro [11]	7B	0.3B	1567	-	79.2	41.0	<u>50.0</u>	-	-
Tar [23]	7B	0.4B	1571	<u>1926</u>	74.4	39.0	-	-	-
Show-o2 [76]	7B	0.5B	<u>1620</u>	-	<u>79.3</u>	**48.9**	-	-	69.8
OneCAT-3B	3B	/	**1630**	**2051**	78.8	41.9	**52.2**	**61.7**	<u>72.5</u>

Table 4 Performance comparison across multiple multimodal understanding benchmarks. Higher scores are better, as indicated by the up-arrow (↑). **A-LLM** denotes the number of activated LLM parameters, while **Vis.** indicates the parameter count of the vision encoder or tokenizer for multimodal understanding. Chameleon [62] does not report the parameter count of its vision tokenizer. slash (/) denotes that models do not require a vision encoder or tokenizer for multimodal understanding. Top-1 accuracy is reported (Best in **bold**, second best is <u>underlined</u>).

strates superior performance, significantly outperforming all existing encoder-free understanding-only MLLMs, e.g., HoVLE [61] and EvEv2 [19], across nearly all benchmarks. For instance, on OCR-related tasks including TextVQA (73.9), ChartQA (81.2), InfoVQA (64.8), and DocVQA (91.2), OneCAT-3B achieves new state-of-the-art results among encoder-free models. It also excels in general vision-language benchmarks such as MME-P (1630), MMBench (78.8), and MM-Vet (52.2).

Moreover, OneCAT-3B outperforms recent unified MLLMs that rely on external vision encoders or tokenizers—such as Janus-Pro-7B [11] (using SigLIP [84]) and Tar-7B [23] (using SigLip2 [64])—despite activating fewer parameters. Compared to top-tier encoder-based understanding-only models like Qwen2.5-VL-3B [3], our model exhibits a slight performance gap, which we primarily attribute to differences in the scale and quality of training data. Specifically, Qwen2.5-VL was trained on 4T tokens, whereas our OneCAT was trained on only 0.5T tokens for multimodal understanding. We believe this gap can be bridged in the future by scaling up the pretraining data and incorporating more higher-quality instruction data.

7.2 Visual Generation.

We evaluate our model on three widely-used visual generation benchmarks: two for text-to-image generation, GenEval [21] and DPG-Bench [27], and one for instruction-based image editing, ImgEdit [78]. We follow previous works [10, 16, 41] to use Classifier-free guidance(CFG) [25] to enhance visual generation quality. During training, we randomly drop tokens of conditional text and reference image with probabilities 0.1, 0.1, respectively. During inference, we combines conditional and unconditional predicted logits to produce outputs that better adhere to the given conditions. To ensure a fair comparison, we adhere strictly to the original raw prompts for the GenEval benchmark, unlike some previous works that employ LLM-based prompt rewriting to enhance performance. As shown in Tab.5, 6, and 7, our OneCAT-3B model demonstrates highly competitive

Figure 6 Text-to-Image comparison.

performance across all tasks.

On GenEval (Tab. 5), which evaluates fine-grained instruction following over object counts, colors, and spatial relationships, OneCAT-3B achieves a SOTA overall score of 0.90. This performance surpasses most all unified

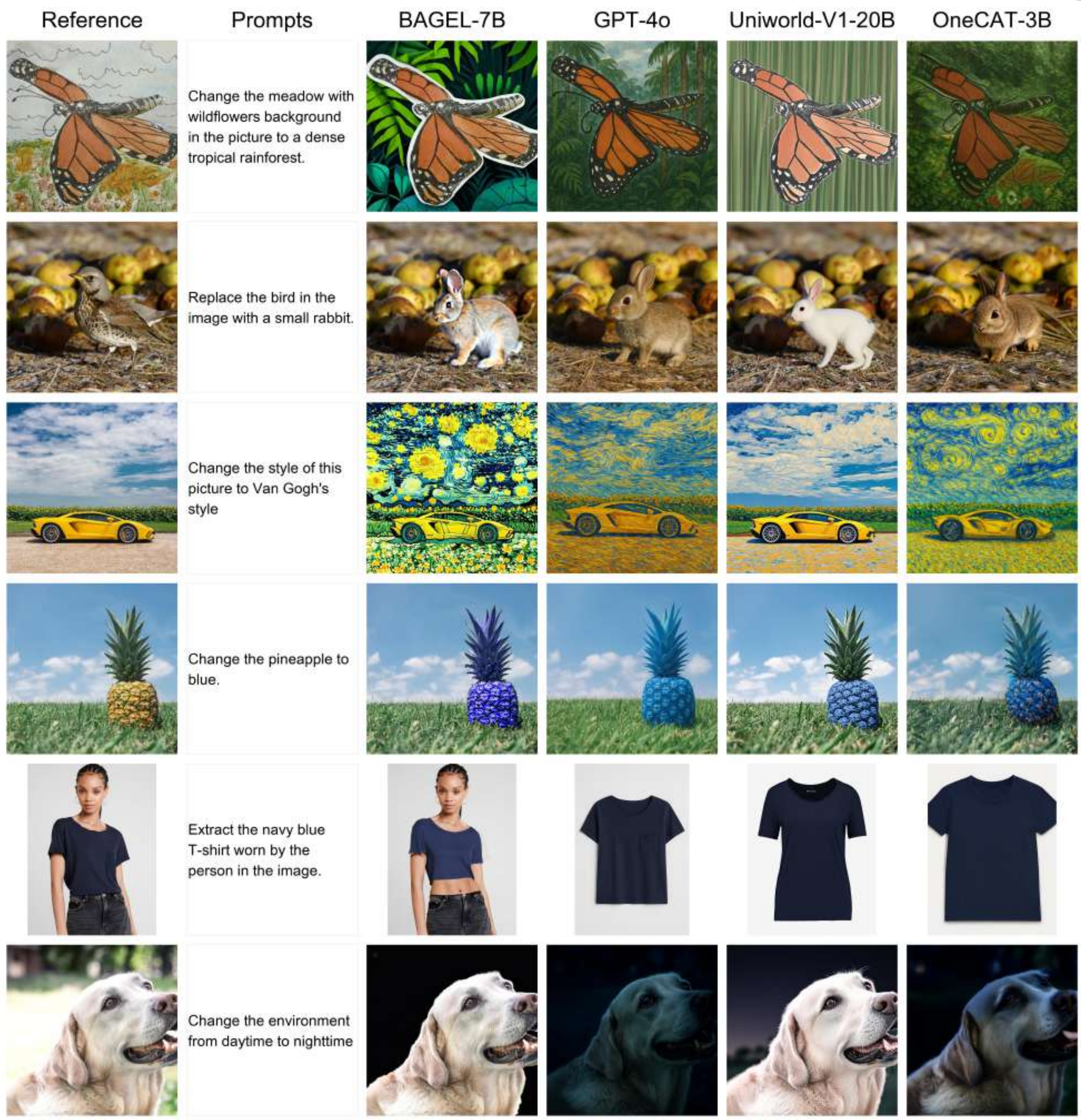

Figure 7 Image-Editing comparison.

models, including the strong baseline BAGEL-7B (0.88 with prompt rewriting) and Mogao-7B (0.89 with prompt rewriting). Notably, OneCAT-3B excels in challenging categories such as Position and Color Attribute, where it achieves the best performance (0.84 and 0.80 respectively), showcasing its superior ability to interpret complex spatial and attribute-based instructions.

On DPG-Bench (Tab. 6), a benchmark focused on compositional text-to-image generation, OneCAT-3B attains a strong overall score of 84.53. This result is highly competitive among unified models, outperforming strong counterparts like Janus-Pro-7B (84.19) and Mogao-7B (84.33).

On ImgEdit-Bench (Tab. 7), a challenging and diverse image editing benchmark, OneCAT-3B achieves an overall score of 3.43. This places it firmly among the top-performing unified models and significantly outperforms many specialized editing models. OneCAT-3B demonstrates exceptional capabilities in categories

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Generation-only Models							
SDXL [54]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALL-E 3 [59]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
Infinity† [24]	-	0.85	-	-	0.49	0.57	0.73
SD3-Medium [20]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev† [34]	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Unified Models							
Chameleon-7B [62]	-	-	-	-	-	-	0.39
Transfusion-7B [90]	-	-	-	-	-	-	0.63
Emu3-8B† [67]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
ILLUME+ 3B [28]	0.99	0.88	0.62	0.84	0.42	0.53	0.72
Harmon-1.5B [71]	0.99	0.86	0.66	0.85	0.74	0.48	0.76
Show-o2-7B [76]	1.00	0.87	0.58	0.92	0.52	0.62	0.76
Janus-Pro-7B [11]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Mogao-7B† [41]	**1.00**	**0.97**	0.83	0.93	**0.84**	**0.80**	<u>0.89</u>
BLIP3-o-8B† [8]	-	-	-	-	-	-	0.84
Tar-7B [23]	0.99	0.92	0.83	0.85	0.80	0.65	0.84
UniWorld-V1-20B [43]	0.99	0.93	0.79	0.89	0.49	0.70	0.80
UniWorld-V1-20B† [43]	0.98	0.93	0.81	0.89	0.74	0.71	0.84
BAGEL-7B [16]	0.99	0.94	0.81	0.88	0.64	0.63	0.82
BAGEL-7B† [16]	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88
OneCAT-1.5B	0.99	0.92	0.83	0.91	0.72	0.75	0.85
OneCAT-3B	**1.00**	<u>0.96</u>	**0.84**	<u>0.94</u>	**0.84**	**0.80**	**0.90**

Table 5 Performance comparison on the GenEval [21] benchmark. The dagger (†) indicates methods that employ an LLM for prompt rewriting. Best in **bold**, second best is <u>underlined</u>.

requiring precise local and global adjustments, securing the top scores in Adjust (3.70), Extract (2.42), and Background (3.79) manipulation. This highlights the effectiveness of our model's ability to condition its generation on fine-grained visual cues from a reference image.

Figures 6 and 7 present qualitative comparisons for the text-to-image and image-editing tasks, respectively. OneCAT-3B exhibits leading instruction-following and world-understanding capabilities among open-source models. For example, in the fourth row of Figure 6, it is the only open-source model that correctly generates exactly four characters in a pixel-art style; in contrast, BAGEL-7B produces five characters, and Janus-Pro-7B fails to render pixel art. Similarly, in the last row of Figure 7, only OneCAT-3B produces an image with the correct lighting conditions.

In summary, the strong instruction-following capability in both text-to-image generation and image editing tasks highlights the effectiveness and versatility of our architectural design.

7.3 Comparison of Inference Efficiency

In addition to its strong performance, our model's architectural design also yields significant improvements in inference efficiency for handling both high-resolution image input and output. We evaluated the efficiency of OneCAT-3B in two distinct phases: prefilling and generation.

In the prefilling phase, we measure the Time to the First Token (TTFT) to assess the computational cost of processing the input. As shown in Tab. 8, thanks to our pure decoder-only architecture that eliminates the need for a separate ViT encoder, OneCAT-3B demonstrates a substantial efficiency advantage, particularly when handling high-resolution images. Compared to QwenVL-3B, as the image resolution increases from 768×768 to 1792×1792, the reduction in TTFT for OneCAT-3B grows sharply from 50.4% to 61.4%. This provides strong evidence for the effectiveness of our design, especially for high-resolution inputs.

In the generation phase, we evaluated the total inference time for text-to-image (T2I) generation and image editing. As detailed in Tab. 9, OneCAT-3B's inference speed far surpasses that of the strong diffusion-based baseline, BAGEL-7B. For instance, when generating a 1024×1024 resolution image, OneCAT-3B's T2I and editing inference times are merely 2.85s and 4.61s, respectively, making it approximately $10\times$ faster than BAGEL-7B. This significant advantage stems from our pioneering multi-scale autoregressive generation

Model	Global	Entity	Attribute	Relation	Other	Overall↑
Generation-only Models						
Hunyuan-DiT [40]	84.59	80.59	88.01	74.36	86.41	78.87
Playground v2.5 [36]	83.06	82.59	81.20	84.08	83.50	75.47
PixArt-Σ [9]	86.89	82.89	88.94	86.59	87.68	80.54
DALL-E 3 [59]	90.97	89.61	88.39	90.58	89.83	83.50
Infinity [24]	93.11	-	-	90.76	-	83.46
SD3-Medium [20]	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev [34]	82.10	89.50	88.80	91.10	89.40	84.00
Unified Models						
Emu3-8B [67]	-	-	-	-	-	81.60
Janus-Pro-7B [11]	86.90	88.90	89.40	89.32	89.48	84.19
Mogao-7B [41]	82.37	90.03	88.26	93.18	85.40	84.33
BLIP3-o-8B [8]	-	-	-	-	-	81.60
Tar-7B [23]	83.98	88.62	88.05	**93.98**	84.86	84.19
Show-o2-7B [76]	89.00	**91.78**	**89.96**	91.81	**91.64**	**86.14**
OneCAT-1.5B	**90.48**	86.70	86.75	89.32	84.93	81.72
OneCAT-3B	85.46	90.81	89.00	90.40	89.56	84.53

Table 6 Performance comparison on the DPG-Bench [27] benchmark. Best in **bold**, second best is underlined.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall
Editing-only Models										
MagicBrush [85]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix [5]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [32]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [87]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
Step1X-Edit [46]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
ICEdit [86]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Unified Models										
OmniGen [73]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
OmniGen2 [70]	3.57	3.06	1.77	**3.74**	3.20	3.57	**4.81**	2.52	**4.68**	**3.44**
BAGEL-7B [16]	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1-20B [43]	**3.82**	3.64	2.27	3.47	**3.24**	2.99	4.21	**2.96**	2.74	3.26
OneCAT-3B	3.65	**3.70**	**2.42**	3.92	3.00	**3.79**	4.61	2.23	3.53	3.43

Table 7 Comprehensive comparison on ImgEdit-Bench [78] showing performance across nine editing categories. Higher scores are better for all metrics. Best in **bold**, second best is underlined.

mechanism inside LLM that enables parallel token generation, which drastically reduces the number of required decoding steps compared to iterative diffusion models.

In summary, OneCAT achieves exceptional inference efficiency in both the prefilling and generation stages. This highlights the immense potential and practical value of our unified autoregressive framework for building efficient and powerful large multimodal models.

8 Conclusion

In this work, we presented OneCAT, a pure decoder-only unified multimodal model that seamlessly integrates understanding, generation, and editing within a single, streamlined architecture. By eliminating external encoders and tokenizers, employing a modality-specific MoE design, and introducing a multi-scale autoregressive generation mechanism, OneCAT achieves strong performance across a wide range of benchmarks while significantly improving inference efficiency. Our results demonstrate the viability and advantages of a first-principles approach to multimodal modeling, offering a powerful new baseline for future research and applications in general-purpose multimodal intelligence.

Model	Resolution of Input Image	#Input Text Tokens	#Input Visual Tokens	TTFT(s)	Reduction
Qwen2.5-VL-3B	768 × 768	24	731	0.135	50.4%
OneCAT-3B	768 × 768	24	731 +256*	0.067	
Qwen2.5-VL-3B	1024 × 1024	24	1395	0.216	57.4%
OneCAT-3B	1024 × 1024	24	1395 +256*	0.092	
Qwen2.5-VL-3B	1792 × 1792	24	4098	0.583	61.4%
OneCAT-3B	1792 × 1792	24	4098 +256*	0.225	

Table 8 Efficiency comparison of OneCAT-3B and QwenVL-3B. Models are tested based on one NVIDIA H800 GPU. We report the time to the first token (TTFT) to measure the reduction of computational cost of prefilling phase. 256* denotes the number of visual tokens for thumbnail.

Model	Resolution of Generated Image	T2I Infer. Time (s)	Edit Infer. Time (s)
BAGEL-7B	512 × 512	8.76	13.45
OneCAT-3B	512 × 512	1.40	2.03
BAGEL-7B	1024 × 1024	26.29	46.44
OneCAT-3B	1024 × 1024	2.85	4.61

Table 9 Generation efficiency comparison of OneCAT-3B and BAGEL. Models are tested based on one NVIDIA H800 GPU. We report total inference time to measure the Text-to-Image(T2I) and Image-Editing efficiency.

References

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

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[4] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

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

[6] 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.

[7] 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, pages 3558–3568, 2021.

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

[9] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In European Conference on Computer Vision, pages 74–91. Springer, 2024.

[10] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. arXiv preprint arXiv:2506.18095, 2025.

[11] 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, 2025.

[12] 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, 2024.

[13] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821, 2024.

[14] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[15] Tongyi Data. Sa1b-dense-caption dataset, 2024. URL https://www.modelscope.cn/datasets/Tongyi-DataEngine/SA1B-Dense-Caption.

[16] 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.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[18] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. Advances in Neural Information Processing Systems, 37:52545–52567, 2024.

[19] Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. arXiv preprint arXiv:2502.06788, 2025.

[20] 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.

[21] 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.

[22] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. arXiv preprint arXiv:2412.05237, 2024.

[23] 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, 2025.

[24] 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 Computer Vision and Pattern Recognition Conference, pages 15733–15744, 2025.

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.

[27] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135, 2024.

[28] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. arXiv preprint arXiv:2504.01934, 2025.

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

[30] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. arXiv preprint arXiv:2404.09990, 2024.

[31] isidentical. moondream2-coyo-5m-captions dataset, 2024. URL https://hf-mirror.com/datasets/isidentical/moondream2-coyo-5M-captions.

[32] Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. arXiv preprint arXiv:2502.05628, 2025.

[33] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. pages 235–251, 2016.

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

[35] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. arxiv:2307.16125, 2023.

[36] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024.

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

[38] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? arXiv preprint arXiv:2406.08478, 2024.

[39] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26763–26773, 2024.

[40] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748, 2024.

[41] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472, 2025.

[42] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472, 2025.

[43] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147, 2025.

[44] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147, 2025.

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2023.

[46] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025.

[47] 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, pages 216–233. Springer, 2024.

[48] 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.

[49] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 24960–24971, 2025.

[50] 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:2203.10244, 2022.

[51] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In WACV, pages 2200–2209, 2021.

[52] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In WACV, pages 1697–1706, 2022.

[53] 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.

[54] 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.

[55] PyTorch Team. FlexAttention: The flexibility of pytorch with the performance of flashattention. PyTorch Blog, 2024. URL https://pytorch.org/blog/flexattention/. Accessed: 2024-09-03.

[56] 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, pages 8748–8763. PMLR, 2021.

[57] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.

[58] 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), pages 2556–2565, 2018.

[59] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. arXiv preprint arXiv:2006.11807, 2020.

[60] 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, pages 8317–8326, 2019.

[61] Chenxin Tao, Shiqian Su, Xizhou Zhu, Chenyu Zhang, Zhe Chen, Jiawen Liu, Wenhai Wang, Lewei Lu, Gao Huang, Yu Qiao, et al. Hovle: Unleashing the power of monolithic vision-language models with holistic vision-language embedding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 14559–14569, 2025.

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

[63] 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.

[64] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

[65] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

[66] 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, 2025.

[67] 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.

[68] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In The Thirteenth International Conference on Learning Representations, 2024.

[69] 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, pages 12966–12977, 2025.

[70] 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, 2025.

[71] 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, 2025.

[72] 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.

[73] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 13294–13304, 2025.

[74] Chunyu Xie, Heng Cai, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, et al. Ccmb: A large-scale chinese cross-modal benchmark. In Proceedings of the 31st ACM International Conference on Multimedia, pages 4219–4227, 2023.

[75] 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.

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

[77] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

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

[79] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. National Science Review, 11(12):nwae403, 2024.

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

[81] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14022–14032, 2024.

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

[83] 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9556–9567, 2024.

[84] 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, pages 11975–11986, 2023.

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

[86] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690, 2025.

[87] Haozhe Zhao, Xiaojian Shawn 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. Advances in Neural Information Processing Systems, 37:3058–3093, 2024.

[88] Haozhe Zhao, Xiaojian Shawn 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. Advances in Neural Information Processing Systems, 37:3058–3093, 2024.

[89] Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. arXiv preprint arXiv:2406.07548, 2024.

[90] 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.

[91] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

[92] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. arXiv preprint arXiv:2501.12327, 2025.