APPENDIX: ONECAT: DECODER-ONLY AUTO-REGRESSIVE MODEL FOR UNIFIED UNDERSTANDING AND GENERATION

Contents

A. Related Work

A.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 (*i.e.,* CLIP [64], SigLIP [91], and InternViT [17]), to a powerful LLM through a trainable connector. Pioneering works [3, 42] propose sophisticated connector designs. For example, Flamingo [3] introduces gated cross-attention layers to inject visual information into an LLM, while BLIP-2 [42] develops the Q-Former to bridge the modality gap between the vision encoder and LLM. A significant shift occurs with LLaVA [50], which simplifies the connector to a lightweight MLP, which become a foundational blueprint for subsequent MLLMs. For example, recent state-of-the-art models like the InternVL series [15–17, 78, 94] and the Qwen-VL series [4, 5, 77] adopt this 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, especially for the **prefilling** stage. Second, the connector 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.

A.2. Decoder-only MLLMs

Decoder-only MLLMs, also known as monolithic MLLMs, have recently emerged as a minimalist yet powerful alternative to the compositional MLLMs. This paradigm aims to achieve greater efficiency and a more direct **early fusion** of modalities by removing the separate vision encoder or tokenizer. For example, Fuyu-8B [6] processes vision patches by feeding them through a simple linear patch embedding layer directly into the LLM, which markedly reduce inference latency. Inspired by this success, subsequent works [21, 22, 37, 53, 68, 76], further advance decode-only MLLMs by targeting their training processes and architectures. EvE [21] and VoRA [76] aligns the LLM's hidden states with semantic features from a pre-trained ViT. However, directly using a smaller model (e.g., a ViT with hundreds of millions of parameters) as the teacher to distill

knowledge into a significantly larger LLM (with several billion parameters) may restrict the expressive capacity of the LLM. Differently, Mono-InternVL [53] and EvEv2.0 [22] adopt a Mixture-of-Experts (MoE) framework, introducing a dedicated *visual expert* to handle visual-specific features more effectively. HoLVE [73] 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 decoder-only MLLMs 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.

A.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 [11, 14, 19, 41, 45, 47, 62, 79, 81, 83, 84, 93]. Pioneering unified MLLMs such as Chameleon [74],Transfusion [93], emu3 [79], show-o [83] and Synergen-VL [41] utilize vison tokenizer (e.g., VQ-VAE) to convert images into 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 [14, 81] 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 [62], BLIP3-O [11], Uniworld-V [47] assembles off-the-shelf specialized MLMMs and diffusion models by tuning adapters and learnable query tokens, which sacrifices true architectural unification for modularity. BAGEL [19] and Mogao [45] 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.

A.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 [75] 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 [26] 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 [95] stack the transformer from pretrained VAR [75] 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 decoder-only transformer of the LLM**, and proposes the scale-aware adapter to further exploit the scale-specific representation for enhanced visual generation.

B. Preliminary of Next-Scale Prediction

B.1. Multiscale Tokenization

Leveraging the inherent coarse-to-fine structure of natural images, VAR [75] introduces a multi-scale tokenizer that encodes an image into K token scales $(R_1, R_2, \ldots, R_K)$. The resolution of each scale R_k, denoted as (h_k, w_k), increases monotonically with the scale index k. Specifically, given a feature map F extracted from an image with an image encoder, VAR defines these token scales recursively:

$$R_1 = \mathcal{Q}(\text{interpolate}_1(F)), \tag{1}$$

$$R_2 = \mathcal{Q}(\text{interpolate}_2(F - \text{interpolate}_K(R_1))), \tag{2}$$

$$\vdots$$

$$R_k = \mathcal{Q}\Big(\text{interpolate}_k(F - \sum_{i=1}^{k-1} \text{interpolate}_K(R_i))\Big), \tag{3}$$

$$\vdots$$

$$R_K = \mathcal{Q}\Big(F - \sum_{i=1}^{K-1} \text{interpolate}_K(R_i)\Big), \tag{4}$$

where interpolate_i is an operator that resizes its input to the resolution (h_i, w_i), and $\mathcal{Q}$ is the quantization operator. For a given 3D feature map $x \in \mathbb{R}^{d \times h \times w}$, we implement $\mathcal{Q}$ using Binary Spherical Quantization (BSQ) [92], following Han

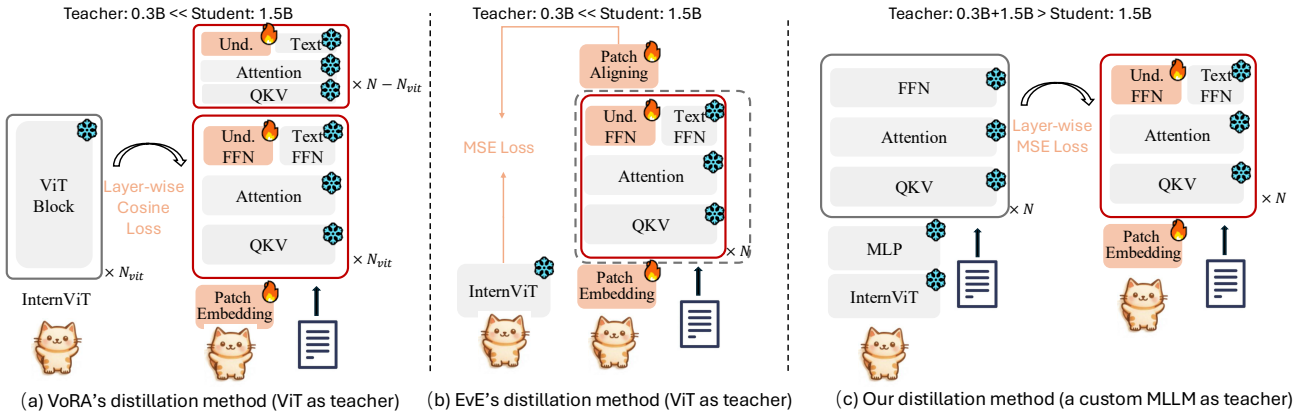


Figure A1. Comparison of existing decoder-only MLLM understanding distillation methods, *i.e.,* VoRA [76], EvE [21], when they are applied to our OneCAT. We omit the `Gen. FFN` in this figure for clarity.

et al. [26]. The quantization is applied to each spatial feature vector $x_{ij} \in \mathbb{R}^d$ as:

$$\mathcal{Q}(x_{ij}) = \frac{1}{\sqrt{d}} \operatorname{sign}\left(\frac{x_{ij}}{\|x_{ij}\|_2}\right). \tag{5}$$

B.2. Visual Auto-Regressive Training

The premise for generation is that the feature map F can be well approximated by summing all scales upsampled to the final resolution: $F \approx \sum_{i=1}^{K} \text{interpolate}_K(R_i)$. It therefore suffices to generate the sequence of scales $R_{1:K}$ to synthesize an image. To achieve this, VAR models the joint distribution over the scales auto-regressively, factorizing the log-likelihood as:

$$\log p_\theta(R_{1:K}) = \sum_{k=1}^{K} \log p_\theta(R_k \mid R_{1:k-1}). \tag{6}$$

The model, with parameters θ, is trained to maximize this log-likelihood by learning to predict the current scale R_k conditioned on all preceding scales $R_{1:k-1}$. To enable efficient parallel decoding, VAR assumes that all tokens within the current scale R_k are conditionally independent given $R_{1:k-1}$.

However, this conditional independence assumption, coupled with imperfect model training, can lead to error propagation: mistakes in generating early-stage scales $(R_1, \ldots, R_{k-1})$ are amplified when generating subsequent, higher-resolution scales R_k. To mitigate this issue, Han et al. [26] proposed *Bitwise Self-Correction*. This technique involves training the model on corrupted versions of the conditioning scales $R_{1:k-1}$, thereby teaching it to generate the correct R_k even when the preceding scales are imperfect. This robustifies the model against its own generation errors during inference.

B.3. Predefined Scale Schedules

We follows Han et al. [26] and Tian et al. [75] to establish a set of predefined scale schedules, thus ensuring efficient training across images with varying aspect ratios. As detailed in Table A1, for each target aspect ratio r, we define a specific schedule as a sequence of K resolution tuples: $\{(h_1^r, w_1^r), \ldots, (h_K^r, w_K^r)\}$.

These schedules are designed based on two fundamental principles: **1.Aspect Ratio Consistency:** Each tuple (h_k^r, w_k^r) within a schedule maintains an aspect ratio that is approximately equal to the target ratio r, especially at larger scales. **2.Consistent Area Across Scales:** For any given scale level k, the image area, calculated as $h_k^r \times w_k^r$, is kept roughly constant across different aspect ratio schedules. This standardization ensures that the training sequence lengths are similar for various aspect ratios, thereby improving overall training efficiency. During the inference stage, these predefined schedules enable the model to generate high-quality images covering a wide range of aspect ratios.

C. Details of Hyperparameter and Configuration of the Training Recipe Across Stages

We present the detailed hyperparameter and configuration of the training recipe across stages in Tab. A2.

D. Details of Class-Free Guidance

We follow previous works [10, 19] to use CFG for enhanced visual generation quality. For training, we randomly drop tokens of conditional text and reference image with probabilities 0.1. For inference, we combine conditional and unconditional predicted logits to produce outputs.

Specifically, for **text-to-image generation**, the final logits $\mathbf{L}_{\text{final}}$ are computed as a linear combination of the con-

Table A1. Predefined scale schedules $\{(h_1^r, w_1^r), \ldots, (h_K^r, w_K^r)\}$ for different aspect ratios. Following Han et al. [26], OneCAT utilizes $K = 13$ scales to generate the highest resolution image, such as a 1024×1024 image for the 1:1 aspect ratio, while lower-resolution images like 512×512 can be produced by truncating the schedule to $K = 10$.

Aspect Ratio	Resolution	Scale Schedule
1.000 (1:1)	1024×1024	(1,1) (2,2) (4,4) (6,6) (8,8) (12,12) (16,16) (20,20) (24,24) (32,32) (40,40) (48,48) (64,64)
0.800 (4:5)	896×1120	(1,1) (2,2) (3,3) (4,5) (8,10) (12,15) (16,20) (20,25) (24,30) (28,35) (36,45) (44,55) (56,70)
1.250 (5:4)	1120×896	(1,1) (2,2) (3,3) (5,4) (10,8) (15,12) (20,16) (25,20) (30,24) (35,28) (45,36) (55,44) (70,56)
0.750 (3:4)	864×1152	(1,1) (2,2) (3,4) (6,8) (9,12) (12,16) (15,20) (18,24) (21,28) (27,36) (36,48) (45,60) (54,72)
1.333 (4:3)	1152×864	(1,1) (2,2) (4,3) (8,6) (12,9) (16,12) (20,15) (24,18) (28,21) (36,27) (48,36) (60,45) (72,54)
0.666 (2:3)	832×1248	(1,1) (2,2) (2,3) (4,6) (6,9) (10,15) (14,21) (18,27) (22,33) (26,39) (32,48) (42,63) (52,78)
1.500 (3:2)	1248×832	(1,1) (2,2) (3,2) (6,4) (9,6) (15,10) (21,14) (27,18) (33,22) (39,26) (48,32) (63,42) (78,52)
0.571 (4:7)	768×1344	(1,1) (2,2) (3,3) (4,7) (6,11) (8,14) (12,21) (16,28) (20,35) (24,42) (32,56) (40,70) (48,84)
1.750 (7:4)	1344×768	(1,1) (2,2) (3,3) (7,4) (11,6) (14,8) (21,12) (28,16) (35,20) (42,24) (56,32) (70,40) (84,48)
0.500 (1:2)	720×1440	(1,1) (2,2) (2,4) (3,6) (5,10) (8,16) (11,22) (15,30) (19,38) (23,46) (30,60) (37,74) (45,90)
2.000 (2:1)	1440×720	(1,1) (2,2) (4,2) (6,3) (10,5) (16,8) (22,11) (30,15) (38,19) (46,23) (60,30) (74,37) (90,45)
0.400 (2:5)	640×1600	(1,1) (2,2) (2,5) (4,10) (6,15) (8,20) (10,25) (12,30) (16,40) (20,50) (26,65) (32,80) (40,100)
2.500 (5:2)	1600×640	(1,1) (2,2) (5,2) (10,4) (15,6) (20,8) (25,10) (30,12) (40,16) (50,20) (65,26) (80,32) (100,40)
0.333 (1:3)	592×1776	(1,1) (2,2) (2,6) (3,9) (5,15) (7,21) (9,27) (12,36) (15,45) (18,54) (24,72) (30,90) (37,111)
3.000 (3:1)	1776×592	(1,1) (2,2) (6,2) (9,3) (15,5) (21,7) (27,9) (36,12) (45,15) (54,18) (72,24) (90,30) (111,37)

Table A2. Detailed hyperparameter and configuration of the training recipe across stages

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∼1776)
Number of Scales : Gen.	-	7	10	10∼13

ditional logits $\mathbf{L}_t$ (with text input) and unconditional logits $\mathbf{L}_\emptyset$ (without text input):

$$\mathbf{L}_{\text{final}} = \lambda_t \cdot \mathbf{L}_t + (1 - \lambda_t) \cdot \mathbf{L}_\emptyset \tag{7}$$

where λ_t is the text guidance scale controlling the influence of the text condition.

For **image editing** tasks, which involve both textual and reference image conditions, we employ a dual-guidance mechanism. Let $\mathbf{L}_{t,i}$ denote the logits with both text and reference image conditions, $\mathbf{L}_t$ the logits with text condition only, and $\mathbf{L}_\emptyset$ the logits without any conditions. The refined logits $\mathbf{L}_c$ are first obtained by blending $\mathbf{L}_{t,i}$ and $\mathbf{L}_t$ using a reference image guidance scale λ_i:

$$\mathbf{L}_c = \frac{\mathbf{L}_{t,i} + \lambda_i \cdot \mathbf{L}_t}{1 + \lambda_i} \tag{8}$$

Then, the final output logits $\mathbf{L}_{\text{final}}$ are computed by combining $\mathbf{L}_c$ with the fully unconditional logits $\mathbf{L}_\emptyset$ using the text guidance scale λ_t:

$$\mathbf{L}_{\text{final}} = \mathbf{L}_\emptyset + \lambda_t \cdot (\mathbf{L}_c - \mathbf{L}_\emptyset) \tag{9}$$

This approach allows flexible control over the influence

Table A3. Effect of CFG for T2I generation on OneCAT-3B

CFG	GenEval	DPG
5	0.87	83.69
10	0.89	84.37
15	0.88	84.59
20	0.90	84.53
25	0.88	84.42

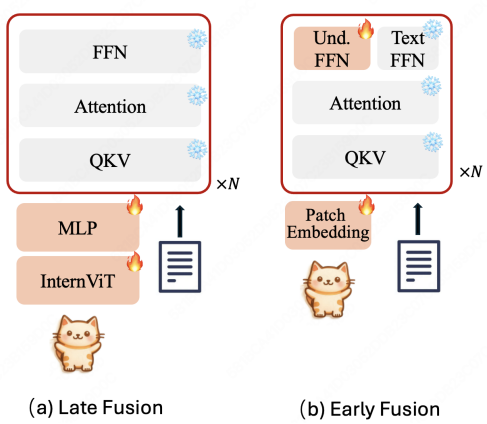


Figure A2. Comparison of late fusion and early fusion for multimodal understanding in our setting. We show the trained modules of pretraining stage and omit the `Gen. FFN` in this figure for clarity.

of both textual and visual conditions during the image editing process.

In our experiments, for text-to-image generation, For text-to-image generation, we set a higher text CFG value of $\lambda_t = 20$. We find that in next-scale-prediction for high-resolution generation (1024x), the number of conditional text tokens is substantially smaller than the vast quantity of image tokens being generated. Therefore, a higher CFG scale is essential to amplify the influence of the limited text condition and prevent it from being diluted during the generation process (as shown in Table A3). For image editing, the reference image itself provides a rich set of visual tokens that serve as a strong guiding signal, so we employ lower CFG values $\lambda_i = 1$ and $\lambda_t = 3$.

E. Additional Ablation Studies

E.1. Effect of Early and Late Fusion in MLLM.

We conduct an ablation study to evaluate the effectiveness and scaling ability of the decoder-only architecture employed in our model compared with the late fusion pipeline for multimodal understanding.

As illustrated in Fig. A2, late fusion corresponds to the conventional encoder-based MLLM approach, *i.e.*, Qwen2.5-VL [77], where images are first processed by a vision encoder before being fed into the LLM. In contrast, early fusion represents the decoder-only MLLM paradigm of our proposed OneCAT.

For late-fusion pretraining, we follow Qwen2.5-VL [5] by employing a randomly initialized vision encoder (InternViT in our experiments), where only the vision encoder and MLP connector are optimized. For early-fusion pretraining, only the patch embedding layer and the understanding expert are optimzied. In both architectures, the LLM is initialized from the pretrained Qwen2.5-1.5B-instruct and remains frozen. Subsequently, all model variants undergo a simplified supervised fine-tuning (SFT) stage on the LLaVA-665k dataset.

As illustrated in Fig. A3 and the corresponding data in Tab. A4, which compare the scaling properties of both models under varying pretraining token budgets, the early and late fusion approaches evince comparable performance. This finding further indicates that the performance gap between OneCAT and Qwen2.5-VL primarily stems from differences in data quality and quantity, rather than the fusion strategy.

E.2. Effect of Distilling Only Visual Tokens

We use the same setting of Sec.4.3.1 in the manuscript to conduct an ablation study to evaluate the impact of distilling different types of tokens. Tab. A5 shows that distilling only the continuous visual tokens results in a slight overall performance drop, suggesting that it is crucial to distill both visual and text tokens.

E.3. Effect on the Increase of Training Tokens of Unified Mid-Training (Stage-2)

Fig. A4 provides the performance of our OneCAT-1.5B model on multimodal understanding and generation benchmarks at various checkpoints throughout the unified mid-training stage, corresponding to different amounts of training tokens.

Before downstream evaluation, the model of each checkpoint undergoes a simplified unified SFT with a combined instruction dataset (LLaVA-665K for understanding and BLIP3o-60K for generation) , and we oversample the BLIP3o-60K dataset to achieve a 1:1 training token ratio for understanding and generation tasks.

E.4. Effect of Multi-Stage Training Pipeline

Tab. A6 provides the performance of OneCAT-1.5B across different training stage. Before downstream evaluation, the model of Stage-1 and Stage-2 also undergoes a simplified unified SFT with a combined instruction dataset (LLaVA-665K and BLIP3o-60K).

Table A4. Performance comparison of different fusion strategies and distillation methods for multimodal understanding across varying training scales in pretraining (stage-1).

Methods	#Trained Tokens of Stage-1	MMB	MME-S	MMVet	SEED	AI2D	ChartQA	TextVQA	**Avg.**
Encoder-based Late Fusion	8B	43.0	1222	13.3	46.8	52.1	10.1	10.6	31.4
	20B	49.0	1437	16.7	50.4	54.4	11.6	11.3	35.0
	70B	51.7	1426	19.2	55.4	55.9	12.0	19.7	37.8
Decoder-only Early Fusion	8B	42.0	1209	13.2	47.0	53.8	10.5	10.0	31.4
	20B	45.7	1312	16.7	51.8	56.4	11.7	11.9	34.4
	70B	50.9	1423	16.9	57.4	57.2	14.2	21.0	38.3
Decoder-only Early Fusion + Proposed Distillation	8B	49.4	1327	15.5	56.3	54.4	11.9	11.3	35.3
	20B	54.3	1410	17.7	61.0	55.4	13.6	15.8	38.3
	70B	57.6	1476	23.4	63.0	57.2	15.0	25.0	42.0
	300B	60.7	1526	27.1	63.4	60.0	19.2	35.7	45.8

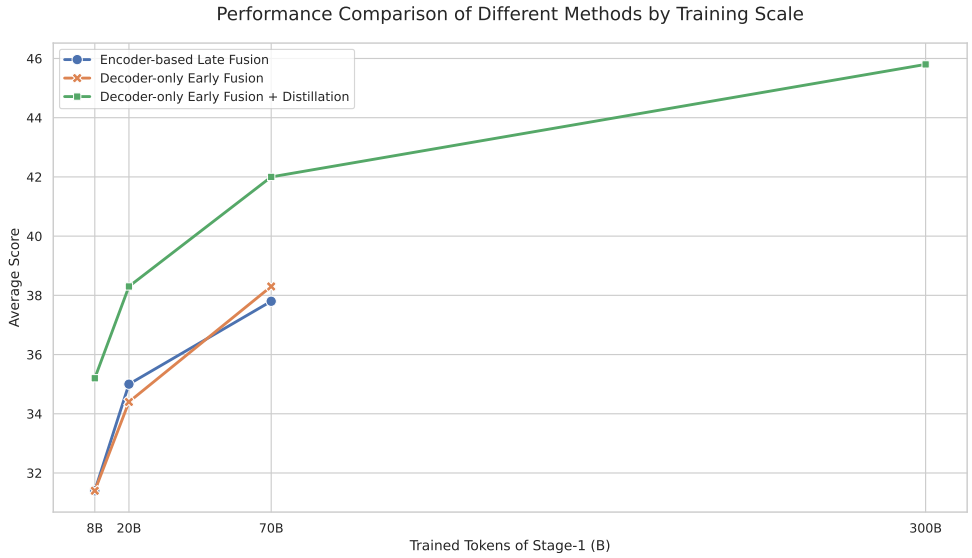


Figure A3. Performance comparison of different methods for multimodal understanding across varying training scales.

Table A5. Effect of distilling only visual tokens

Methods	MMB	MME-S	MMVet	SEED	AI2D	ChartQA	TextVQA	**Avg.**
w/o distillation	42.0	1209	13.2	47.0	53.8	10.5	10.0	31.4
distill only visual tokens	48.1	1299	16.7	55.7	55.2	11.4	10.5	34.8
distill both visual and text tokens	49.4	1327	15.5	56.3	54.4	11.9	11.3	**35.3**

E.5. Effect of Modality-MoE

We conduct an ablation study to evaluate the effectiveness of our proposed Modality-MoE. Three model variants are compared: (i) **Our Modality-MoE**: The structure used in OneCAT, featuring duplicated FFN layers for the understanding, generation, and text experts; (ii) **Modality Mixture-of-Transformers (MoT)**: Following BAGEL [19], this variant duplicates both the FFN and QKV layers across modalities; (iii) **Shared Transformer**: A baseline where the entire transformer block is shared across all modalities, without any modality-specific parameters. Specifically, we train each model using 8B sampled tokens for multimodal understanding pretraining (w/o distillation) and 8B sampled token for generation pretraining, and then undergoes a simplified SFT with a combined instruction dataset (LLaVA-665K and BLIP3o-60K).

Results in Tab. A7 show that both Modality-MoE and MoT surpass the shared transformer by a large margin,

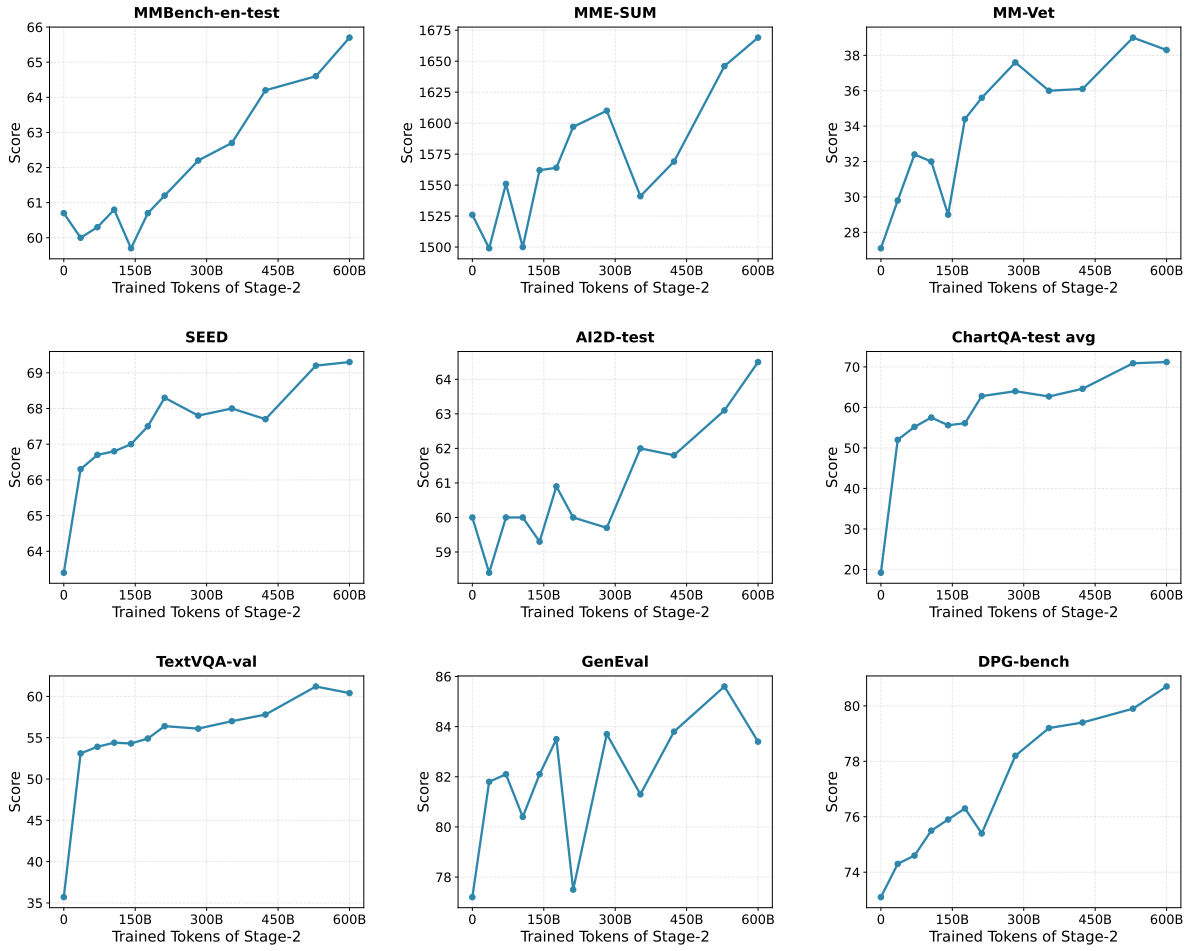


Figure A4. Performance of OneCAT-1.5B on different multimodal understanding and generation benchmarks with the increase of training tokens of unified mid-training (Stage-2).

Table A6. Effect of Multi-Stage Training Pipeline. Before downstream evaluation, the model of Stage-1 and Stage-2 also undergoes a simplified unified SFT with a combined instruction dataset (LLaVA-665K for understanding and BLIP3o-60K for generation).

	Understanding								Generation	
	MMB	MME-S	MMVet	SEED	AI2D	ChartQA	TextVQA	**Avg.**	GenEval	DPG
Stage1-Teacher+ Simplified SFT	64.3	1604	29.1	65.5	59.7	25.2	51.3	50.3	-	-
Stage1 + Simplified SFT	60.7	1526	27.1	63.4	60.0	19.2	35.7	45.8	77.0	73.1
Stage2 + Simplified SFT	65.7	1669	38.3	69.3	64.5	71.2	60.4	61.2	83.4	80.7
Stage3	72.4	1893	42.4	70.9	72.4	76.2	67.0	66.9	85.0	81.7

highlighting the importance of modality-specific parameters. Moreover, our Modality-MoE performs on par with MoT, demonstrating that the MoE design alone is adequate for **the unified autoregressive model**, offering a superior balance of performance and efficiency within a more elegant and simplified architectural framework.

E.6. Additional Efficiency Comparison

As shown in Table A8, we further compare the visual generation efficiency of OneCAT-1.5B/3B with BAGEL-1.5B. It should be noted that the pretrained weights of BAGEL-1.5B are not publicly available; therefore, we conducted in-house training for a sufficient period to obtain a comparable model for this evaluation. The results demonstrate that our OneCAT models achieve a significant speedup while main-

Table A7. Effect of Modality-MoE. We sample 8B tokens for multimodal understanding (w/o distillation) and 8B tokens for generation during the pretraining stage (Stage-1).

	Understanding								Generation	
	MMB	MME-S	MMVet	SEED	AI2D	ChartQA	TextVQA	**Avg.**	GenEval	DPG
Modality MoE	42.0	1209	13.2	47.0	53.8	10.5	10.0	31.4	59.0	62.1
Modality MoT	42.4	1237	13.3	46.6	52.9	10.2	10.4	31.5	59.6	61.7
Shared Transformer	39.9	1162	13.4	44.0	47.2	8.6	9.3	29.1	47.0	53.7

Table A8. Efficiency comparison for generation (**Right**), tested on one NVIDIA H800. **Right:** We report total inference time for Text-to-Image (T2I) and Image-Editing.

Model	**Resolution of Output Image**	**T2I Infer. Time (s)**	**Edit Infer. Time (s)**
BAGEL-1.5B	512 × 512	6.93	9.46
OneCAT-1.5B	512 × 512	1.08 (84%↓)	1.57(83%↓)
OneCAT-3B	512 × 512	1.40 (80%↓)	2.03 (79%↓)
BAGEL-1.5B	1024 × 1024	18.76	27.34
OneCAT-1.5B	1024 × 1024	2.03 (89%↓)	3.46 (87%↓)
OneCAT-3B	1024 × 1024	2.85 (85%↓)	4.61 (83%↓)

taining comparable activated parameters count.

F. Visualization of Discrete Visual Tokens of Different Scales

We generate two 1024× 1024 images and present the visualization of discrete visual tokens across different scales and LLM layers. Inspired by [39, 40], we visualize the intensity of frequency component by applying Fast Fourier Transform (FFT) to the corresponding tokens' feature maps. As shown in Fig A5, the results show that tokens at lower scales primarily encode low-frequency global information, while higher-scale tokens capture high-frequency details, validating the design rationale of our scale-aware adapter.

G. More Qualitative Results

In Fig. A6 and A7, we present qualitative comparisons for text-to-image generation and image editing against several open-source models—including Janus-Pro [14], BAGEL [19], and UniWorld-V1 [47]—as well as the proprietary model GPT-4o-image [61]. We further present additional qualitative results to comprehensively demonstrate the capabilities of our model. Fig. A8 presents text-to-image generation results from OneCAT under various aspect ratios and resolutions. Fig. A9 showcases OneCAT's performance on a range of image editing tasks, such as style transfer, object adjustment, attribute modification, object removal, and background editing. Additionally, Fig. A10 provides examples of OneCAT's multimodal understanding abilities across mathematical reasoning, optical character recognition (OCR), and detailed image captioning.

H. Detailed Benchmark Information

In Tab.2 of the manuscript, the following benchmark abbreviations are used: MMB for MMBench-en-test [51], MME-P for MME-Perception [86], MME-S for MME-Sum [86], MMMU for MMMU-Val [90], MMVet for MM-Vet [89], SEED for Seed-bench [38], MathVista for MathVista-testmini [52], TextVQA for TextVQA-val [69], ChartQA for ChartQA-test [56], InfoVQA for InfoVQA-test [59], DocVQA for DocVQA-test [57], GQA for GQA-testdev [27], and AI2D for AI2D-test [32]. For VQA benchmarks, we compute the average scores of TextVQA, ChartQA, InfoVQA, DocVQA, GQA, and AI2D. For general multimodal benchmarks, the average is computed over MME-S (normalized to a 0–100 scale, where a maximum score of 2800 corresponds to 100), MMB, MMMU, MMVet, MathVista, and SEED.

I. Data Setup Details

We summary the data source of each training stage in Tab. A9.

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 [43], Capsfusion [87], Detailed-Caption [44], SA1B-Dense-Caption [18], and Moondream2-COYO-5M-Captions [31]. (2) Re-captioned Image Datasets: We generate new image-text pairs by re-captioning large-scale public image collections using Qwen2-VL [77]. The source image datasets for this process include COYO700M [7], CC12M [8], CC3M [67], LAION-400M [66], and Zeor250M [82]. From this large-scale dataset, we randomly sample a small-scale 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

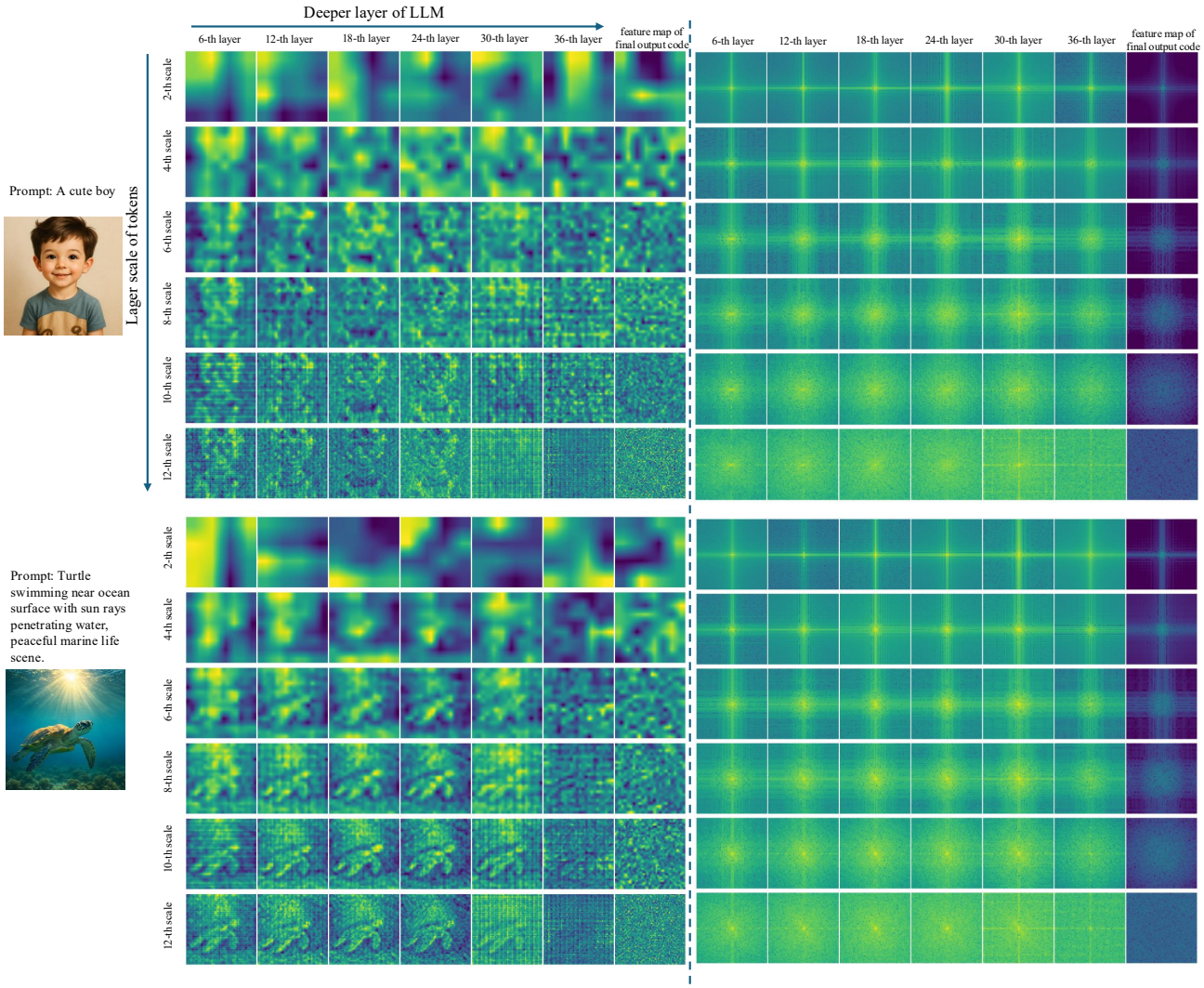


Figure A5. Visualization of discrete visual tokens across scales and LLM layers. **Left:** Each row shows the reshaped feature maps of token hidden states at a specific scale throughout LLM layers. The final column displays the feature map of final output codes fed to the image detokenizer for image reconstruction. All features are resized to 64×64 for display. **Right:** Frequency intensity map of the corresponding feature maps. Lighter colors indicate larger magnitudes, while pixels closer to the center represent lower frequencies. Zoom in better.

to remove samples with low resolution or poor aesthetic scores. This collection consists of 1 million class-labeled images from ImageNet-1k [20], 20 million pairs from public collections (*i.e.,* COYO700M [7], LAION-400M [66] and CC12M [8]), 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 curated 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, which are sourced from Detailed-Caption [44], ALLaVA [9], ShareGPT4V [12], SA1B-Dense-Caption [18], WIT [71], pdfa-eng-wds [1], UReader [85], DVQA [58], OCR-VQA [60], WebSRC [13], GQA [28], visual-genome [34], GRIT [63], and other in-house synthetic visual instruction data.

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 [88], UltraEdit [**?**], HQ-Edit [29] and

Table A9. Summary of Datasets Source in Each Training Stage

Stage	Task Type	Data Sources
Stage-1	Multimodal Understanding	Recap-DataComp-1B [43], Capsfusion [87], Detailed-Caption [44], SA1B-Dense-Caption [18], Moondream2-COYO-5M-Captions [31], COYO700M [7], CC12M [8], CC3M [67], LAION-400M [66], and Zeor250M [82].
	Visual Generation	ImageNet-1k [20], COYO700M [7], LAION-400M [66], CC12M [8], and additional synthetic images generated by FLUX.
Stage-2	Multimodal Understanding & Text-only Instruction	Detailed-Caption [44], ALLaVA [9], ShareGPT4V [12], SA1B-Dense-Caption [18], WIT [71], pdfa-eng-wds [1], UReader [85], DVQA [58], OCR-VQA [60], WebSRC [13], GQA [28], visual-genome [34], GRIT [63], and additional in-house visual and text-only instruction samples.
	Visual Generation	Visual generation data of Stage-1, AnyEdit [88], UltraEdit [**?**], HQ-Edit [29], and OmniEdit [80].
Stage-3	Multimodal Understanding & Text-only Instruction	MAmmoTH-VL [25], AI2D [33], OKVQA [55], VQAv2 [24], ART500K [54], ScienceQA [65], GQA [28], CLEVR-Math [49], COCO-ReM [70], TallyQA [2], Docmatix [36], DVQA [58], DreamSim [23], and ShareGPT4o [16].
	Visual Generation	UniWorld [48], BLIP3o-60k [11], ShareGPT-4o-Image [10], and additional synthetic data generated by GPT-4o and FLUX using the partial prompts from JourneyDB [72].

OmniEdit [80].

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. This dataset comprises 10 million filtered samples from MAmmoTH-VL dataset [25] and 3 million samples from other open-source datasets AI2D [33], OKVQA [55], VQAv2 [24], ART500K [54], ScienceQA [65], GQA [28], CLEVR-Math [49], COCO-ReM [70], TallyQA [2], Docmatix [36], DVQA [58], DreamSim [23], ShareGPT4o [16].

For visual generation, we utilize a total of 3 million samples, aggregated from UniWorld [48], BLIP3o-60k [11], ShareGPT-4o-Image [10], and additional synthetic data generated by GPT-4o [30] and FLUX [35] using the partial prompts from JourneyDB [72]. The overall training token ratio across text-only, multimodal understanding, and visual generation for unified sft is approximately **1 :5 :6**.

J. Other Implementation Details

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 Tab. A2, 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 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 [46], 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.

Figure A6. Text-to-Image comparison.

Figure A7. Image-Editing comparison.

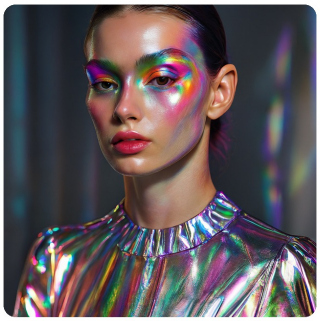
Penguin sliding on ice under aurora lights, comical pose, arctic environment with colorful sky reflection.

Turtle swimming near ocean surface with sun rays penetrating water, peaceful marine life scene.

Van Gogh's Starry Night reimagined with neon cityscape

Sci-fi warrior woman with glowing visor, electric sparks, metallic reflections, futuristic armor.

Quantum portal opening in desert ruins, fractal energy waves, archaeologists in exosuits.



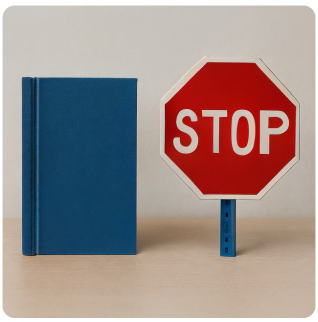
Fashion model with iridescent makeup, prismatic light reflections, high-fashion studio setting

超写实冰川洞穴，蓝冰透射阳光，冰锥如水晶吊灯，地下暗河反光

Charcoal sketch of an old wizard's study, ancient books and potion bottles, dramatic shadows

Film noir detective close-up, venetian blind shadows, cigarette smoke swirls

Magical library pixel scene, floating books, glowing runes, enchanted atmosphere.



a photo of a red stop sign right of a blue book

一个戴帽子的男人，特写镜头，黑白照片，高对比度，面部细节清晰，背景模糊，穿着深色外套，胡须和短发

Figure A8. Showcase of the text-to-image abilities of the OneCAT model.

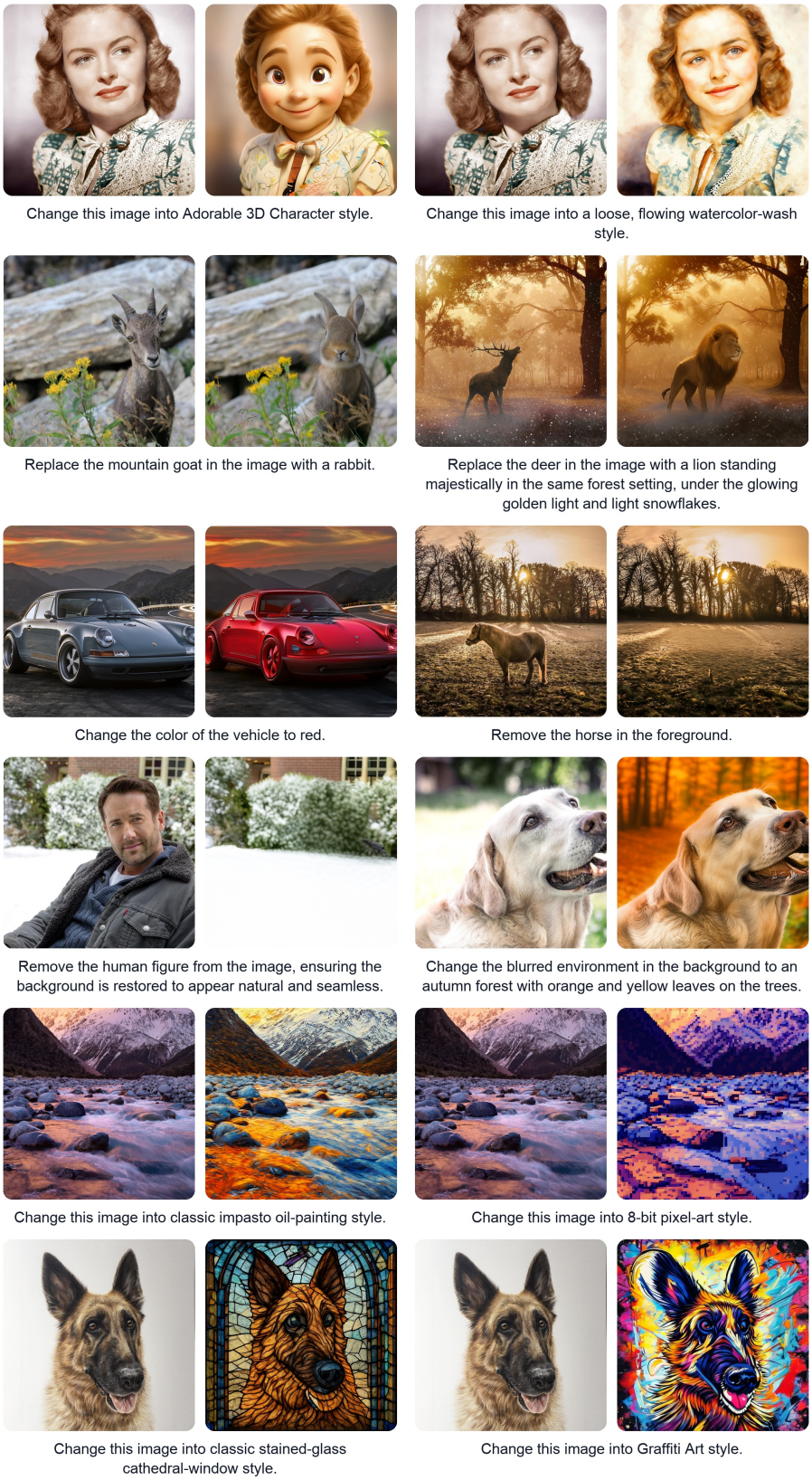


Figure A9. Showcase of the image editing abilities of the OneCAT model.

Solve the following equations:

1) $8x + 11 = 4x + 14$

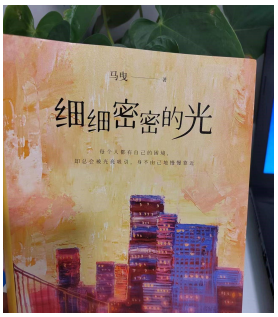
2) $7d - 4 = 11d - 9$

Question:
What is d in the last equation?

Response:
To find the value of d in the last equation, we need to solve the equation 7d - 4 = 11d - 9 .
Step 1: Subtract 7d from both sides of the equation:
7d - 4 - 7d = 11d - 9 - 7d
-4 = 4d - 9
Step 2: Add 9 to both sides of the equation:
-4 + 9 = 4d - 9 + 9
5 = 4d
Step 3: Divide both sides by 4:
\frac{5}{4} = \frac{4d}{4}
d = \frac{5}{4}
Therefore, the value of d is \frac{5}{4}.

Question:
请详细描述该图片。

Response:
这是一张电影海报，背景为一片波涛汹涌的海面，海浪翻滚，水花四溅。海面上有几条巨大的龙，它们张牙舞爪，似乎在进行激烈的战斗。海报中央是一个手持长矛的男孩，他站在海浪中，背对观众，面向龙的方向。男孩的头发随风飘扬，身上穿着红色的腰带，腰带上系着红色的丝带。男孩的长矛顶端有火焰，火焰在海浪中闪烁，显得格外醒目。海报上方有一行红色的文字："魔童归来 风火开年"，表示电影的主题和上映时间。海报下方有电影的标题"哪吒之魔童闹海"，以及导演和上映时间的信息。标题下方的"2025 大年初一"表示电影将在2025年大年初一上映。海报的最下方还有两个标志，一个是"可可豆动画"，另一个是"COLD ROOM"。通过这些信息，可以推断出这是一部以"哪吒"为主题的动画电影，讲述的是哪吒在海中大战巨龙的故事。海报通过生动的海浪和龙的描绘，营造出紧张刺激的氛围，吸引观众的注意力。



Question:
What are all the Chinese scene text in the image?

Response:
马曳 著 细细密密的光 每个人都有自己的困境， 却总会被光亮吸引，身不由己地慢慢靠近

Question:
What is the original price for pork belly before discount?

Response:
To find the original price for pork belly before the discount, we need to use the given information and perform a simple calculation.

1. Identify the Discounted Price and Discount Percentage:
- The discounted price of pork belly is 9.80.
- The discount percentage is 30%.

2. Understand the Relationship Between Original Price, Discount, and Discounted Price:
- The discounted price is the original price minus the discount amount.
- The discount amount is calculated as a percentage of the original price.

3. Calculate the Discount Amount: Let P be the original price.
- The discount amount is 0.30 \times P (since 30% of P is the discount).

4. Set Up the Equation:
- The discounted price is the original price minus the discount amount:[9.80 = P - (0.30 \times P)]

5. Solve for P:
- Combine like terms:[9.80 = P - 0.30P][9.80 = 0.70P]
- Divide both sides by 0.70 to isolate P:[P = \frac{9.80}{0.70}][P = 14]

Therefore, the original price for pork belly before the discount is 14.

Figure A10. Showcase of the visual understanding abilities of the OneCAT model.

References

[1] pdfa-eng-wds. 9, 10

[2] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8076–8084, 2019. 10

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35*, pages 23716–23736, 2022. 1

[4] 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. 1

[5] 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. 1, 5

[6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 1

[7] 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. 8, 9, 10

[8] 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, 9, 10

[9] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024. 9, 10

[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. 3, 10

[11] 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. *arXiv preprint arXiv:2505.09568*, 2025. 2, 10

[12] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 9, 10

[13] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021. 9, 10

[14] 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. 2, 8

[15] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1

[16] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024. 10

[17] 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. Intern VL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 1

[18] Tongyi Data. Sa1b-dense-caption dataset, 2024. 8, 9, 10

[19] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2, 3, 6, 8

[20] 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. 9, 10

[21] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free

vision-language models. In *Advances in Neural Information Processing Systems 37*, pages 52545–52567, 2024. 1, 3

[22] 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. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21014–21025, 2025. 1, 2

[23] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 10

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 10

[25] 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. 10

[26] 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 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15733–15744, 2025. 2, 3, 4

[27] 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. 8

[28] 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. 9, 10

[29] 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. 9, 10

[30] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 10

[31] isidentical. moondream2-coyo-5m-captions dataset, 2024. 8, 10

[32] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. pages 235–251, 2016. 8

[33] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 10

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 9, 10

[35] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 10

[36] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 10

[37] Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single Transformer. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20758–20769, 2025. 1

[38] 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. 8

[39] Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Frequency-aware Transformer for learned image compression. In *The Twelfth International Conference of Learning Representations*, 2024. 8

[40] Han Li, Shaohui Li, Shuangrui Ding, Wenrui Dai, Maida Cao, Chenglin Li, Junni Zou, and Hongkai Xiong. Image compression for machine and human vision with spatial-frequency adaptation. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 382–399, 2024. 8

[41] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29767–29779, 2025. 2

[42] 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. 1

[43] 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. 8, 10

[44] 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. 8, 9, 10

[45] 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. 2

[46] 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. 11

[47] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. UniWorld-V1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 2, 8

[48] 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. 10

[49] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 10

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, pages 34892–34916, 2023. 1

[51] 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. 8

[52] 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. 8

[53] 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 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24960–24971, 2025. 1, 2

[54] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1183–1191, 2017. 10

[55] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 10

[56] 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. 8

[57] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 8

[58] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 9, 10

[59] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 8

[60] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 9, 10

[61] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 8

[62] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with MetaQueries. In *arXiv preprint arXiv:2504.06256*, 2025. 2

[63] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824, 2023. 9, 10

[64] 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. 1

[65] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 10

[66] 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. 8, 9, 10

[67] 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. 8, 10

[68] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951*, 2025. 1

[69] 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. 8

[70] Shweta Singh, Aayan Yadav, Jitesh Jain, Humphrey Shi, Justin Johnson, and Karan Desai. Benchmarking object detectors with coco: A new path forward. In *European Con-*

ference on Computer Vision, pages 279–295. Springer, 2024. 10

[71] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021. 9, 10

[72] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. JourneyDB: A benchmark for generative image understanding. In *Advances in Neural Information Processing Systems 36*, pages 49659–49678, 2023. 10

[73] Chenxin Tao, Shiqian Su, Xizhou Zhu, Chenyu Zhang, Zhe Chen, Jiawen Liu, Wenhai Wang, Lewei Lu, Gao Huang, Yu Qiao, and Jifeng Dai. Hovle: Unleashing the power of monolithic vision-language models with holistic vision-language embedding. In *Proceedings of the 2025 IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, pages 14559–14569, 2025. 2

[74] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[75] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems 37*, pages 84839–84865, 2024. 2, 3

[76] Han Wang, Yongjie Ye, Bingru Li, Yuxiang Nie, Jinghui Lu, Jingqun Tang, Yanjie Wang, and Can Huang. Vision as LoRA. *arXiv preprint arXiv:2503.20680*, 2025. 1, 3

[77] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 5, 8

[78] 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. 1

[79] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2

[80] 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. 10

[81] 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 2025 IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 12966–12977, 2025. 2

[82] 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. 8, 10

[83] 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. 2

[84] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models, 2025. 2

[85] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. 9, 10

[86] 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. 8

[87] 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. 8, 10

[88] 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. 9, 10

[89] 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. 8

[90] 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. 8

[91] 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. 1

[92] Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image

and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024. 2

[93] 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. 2

[94] 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. 1

[95] 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. 2