Auto-Encoding Morph-Tokens for Multimodal LLM

Kaihang Pan [1] **Siliang Tang** [1] **Juncheng Li** [1 2] **Zhaoyu Fan** [1] **Wei Chow** [1] **Shuicheng Yan** [3] **Tat-Seng Chua** [2]
Yueting Zhuang [1] **Hanwang Zhang** [3 4]

Abstract

For multimodal LLMs, the synergy of visual comprehension (textual output) and generation (visual output) presents an ongoing challenge. This is due to a conflicting objective: for comprehension, an MLLM needs to abstract the visuals; for generation, it needs to preserve the visuals as much as possible. Thus, the objective is a dilemma for visual-tokens. To resolve the conflict, we propose encoding images into *morph-tokens* to serve a dual purpose: for comprehension, they act as visual prompts instructing MLLM to generate texts; for generation, they take on a different, non-conflicting role as complete visual-tokens for image reconstruction, where the missing visual cues are recovered by the MLLM. Extensive experiments show that morph-tokens can achieve a new SOTA for multimodal comprehension and generation simultaneously. Our project is available at https://github.com/DCDmllm/MorphTokens.

1. Introduction

State-of-the-art Multimodal Large Language Models (MLLMs) are still facing a great divide between visual comprehension (textual output) and generation (visual output). For comprehension tasks—"Tell me why the image [IMG] is funny"—we use GPT-4V (Achiam et al., 2023); yet for generation—"Turn the image [IMG] into the style of Ghibli"—instead, we need DALL·E (Ramesh et al., 2021). Therefore, the community is interested in a *unified, token-based, auto-regressive* MLLM framework (Yu et al., 2023b). Unlike traditional multimodal large models (Wang et al., 2022a; Rahman et al., 2020), this framework employs LLM as the core reasoning engine that drives both multimodal comprehension and generation.

[1]Zhejiang University [2]National University of Singapore [3]Skywork AI [4]Nanyang Technological University. Correspondence to: Juncheng Li <junchengli@zju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

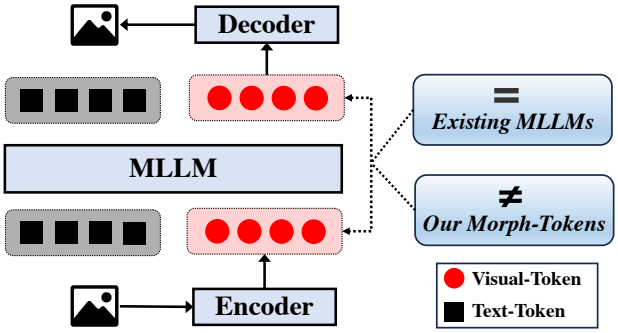

Figure 1. Comparison between existing MLLMs and ours. The key difference is the equality between pre- and post-MLLM visual-tokens in training time.

Existing solutions are straightforward. As shown in Figure 1, they have three steps: 1) images are encoded into visual-tokens by a tokenizer (Fang et al., 2023; Esser et al., 2021); 2) these pre-MLLM visual-tokens are fed into an MLLM to complete vision-language comprehension tasks, where the MLLM is usually initialized from a pre-trained LLM, and 3) the post-MLLM visual-tokens are used to reconstruct input image in training or generate new images such as image editing in testing (Koh et al., 2023; Yu et al., 2023b). More details of such MLLMs are reviewed in Section 2.

Yet, the synergy of comprehension and generation is not achieved. The primary challenge lies in the conflicting MLLM's training objectives for comprehension and generation tasks. Comprehension may discard visual features due to the need for visual abstraction, *i.e.*, the MLLM training encourages the image tokenizer to output pre-MLLM visual-tokens invariant to task-irrelevant visual changes (Dai et al., 2023; Li et al., 2023c)—a many-to-one map from images to tokens; conversely, generation requires preserving the visual details as much as possible, *i.e.*, the post-MLLM visual-tokens should be equivariant to all the visual changes (Wang et al., 2023)—a one-to-one map from tokens to images. The requirement for equality between pre- and post-MLLM visual-tokens poses a dilemma in the auto-regressive training of multimodal token sequences.

As shown in Figure 2(a&b), the comprehension (or generation) performance consistently decreases as the number of generation (or comprehension) training data increases, and vice versa. Another piece of evidence is shown in

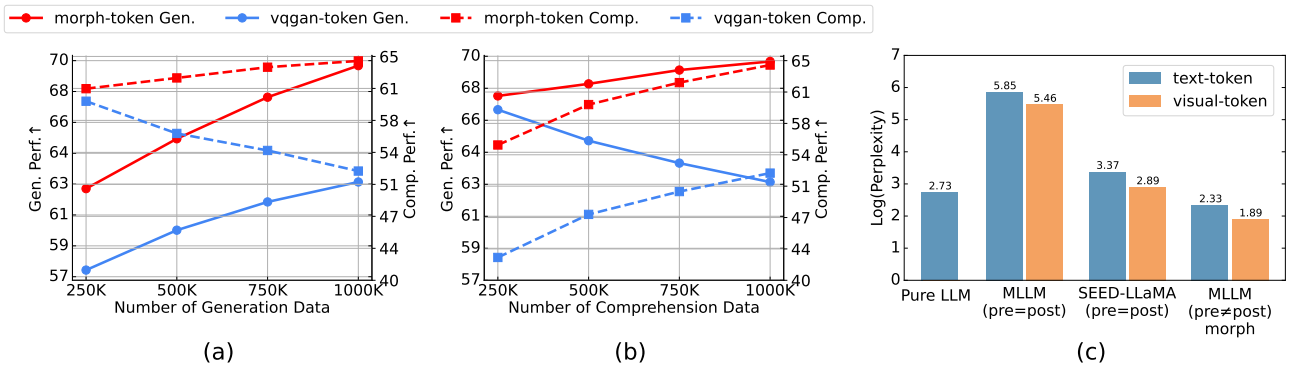

Figure 2. (a, b) When the amount of visual comprehension (or generation) training data is fixed, the trend of comprehension & generation performance with the number of visual generation (or comprehension) training data. (c) The average perplexity of the post-MLLM text-/visual-tokens with 1k text-image pair inputs. "pre/post" denotes pre-/post-MLLM.

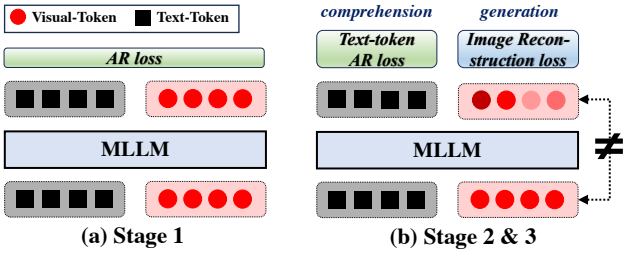

Figure 3. Our 3-stage training strategy (Section 3.4). AR: auto-regressive.

Figure 2(c), if we freeze an image tokenizer that encodes visually complete tokens capable of perfect image reconstruction (*e.g.*, VQ-GAN (Esser et al., 2021)), the multimodal auto-regressive training may even harm the original LLM's performance (Kondratyuk et al., 2023; Yu et al., 2022), that is, the perplexity of both the textual and visual-tokens is significantly increased.

We propose **Morph-Tokens** to resolve the conflict. As illustrated in Figure 1, the term "morph" implies a transformation where the pre-MLLM visual-tokens are *not necessarily equal* to the post-MLLM ones. Specifically, the pre-MLLM tokens are abstract semantics, serving as visual prompts for comprehension tasks. In contrast, the post-MLLM tokens are visually complete tokens for image generation, thanks to the powerful comprehension ability of MLLM that recovers the lost visual features due to abstraction. We will detail the implementation of morph-token-based MLLM in Section 3.2.

How do we use vision-language data to train morph-tokens to achieve the dual purpose without conflict? The key is to detach the textual and image reconstruction losses by using morph-tokens. This approach trains the MLLM to recognize the abstract pre-MLLM visual-tokens for comprehension. Simultaneously, it ensures their recovery back to visually complete tokens for image generation. To this end, we propose a 3-stage training strategy (Section 3.4):

Stage 1: As shown in Figure 3(a), we use image-text pairs to train the morph-token encoder and the MLLM (initialized by an LLM) to auto-regress the concatenated morph-token and text-token sequence. This stage expands the token vocabulary, transitioning from LLM to MLLM. Note that although this stage requires equality between pre- and post-MLLM morph-tokens, there is no conflict due to the absence of a visual generation objective.

Stage 2: As shown in Figure 3(b), we use the same image-text pairs to train the morph-token encoder, MLLM, and the decoder by both comprehension and generation tasks. For comprehension, *i.e.*, image captioning, the pre-MLLM morph-tokens act as visual prompts instructing the MLLM to generate textual captions of the image; for generation, *i.e.*, text-to-image generation, the post-MLLM morph-tokens play a different, non-conflicting role as visually complete tokens to reconstruct the input image. This stage can be viewed as an *auto-encoding* process, unique in that it does not have a fixed morph-token bottleneck.

Stage 3: Similar to Stage 2, we use various vision-language tasks including both comprehension (e.g., VQA) and generation (e.g., image editing) to instruction-tune everything.

Thanks to morph-tokens and the above training strategy, we observe a preliminary synergy shown in Figure 2. Through extensive experiments (Section 4), besides a new SOTA on challenging vision-language benchmarks (*e.g.*, DEMON (Li et al., 2023c)), we further find that our morph-token-based MLLM significantly outperforms others in multi-turn image editing (Figure 5) and in-context learning (Figure 6). Notably, these remarkable abilities to preserve image fidelity while understanding language instructions are rarely observed in prior works.

2. Related Work

Thanks to Figure 1, we summarize relevant MLLMs into Table 1. We first divide them into two groups based on

Models	pre-MLLM visual-token	post-MLLM visual-token	in-context comp.	advanced image-edit	detached loss
Cogview (Ding et al., 2021)	complete	complete	✗	✗	✗
TEAL (Yang et al., 2023)	complete	complete	✗	✗	✗
CM3Leon (Yu et al., 2023b)	complete	complete	✗	-	✗
VideoPoet (Kondratyuk et al., 2023)	complete	complete	✗	✓	✗
Gill (Koh et al., 2023)	abstract	abstract	✗	✗	✗
Emu-I (Sun et al., 2023b)	abstract	abstract	✓	✗	✗
Emu2-Chat (Sun et al., 2023a)	abstract	abstract	✓	✗	✗
DreamLLM (Dong et al., 2023)	abstract	abstract	✓	✗	✗
LaVIT (Jin et al., 2023)	abstract	abstract	✓	✗	✗
Seed-LLaMA (Ge et al., 2023)	abstract	abstract	✓	✗	✗
AnyGPT (Zhan et al., 2024)	abstract	abstract	✓	✗	✗
Mini-Gemini (Li et al., 2024)	abstract	abstract	✓	✗	✗
Morph-Token (ours)	abstract	complete	✓	✓	✓

Table 1. Positioning of existing MLLMs and ours in terms pre-/post-MLLM visual-tokens, complex comprehension/generation capabilities, and if the conflicting training losses is detached.

whether they use VQ-GAN (Esser et al., 2021) or Stable Diffusion (Rombach et al., 2022) as the Decoder. Columns 2&3 denote the nature of pre- and post-MLLM visual tokens, indicating whether they contain abstracted semantics or complete visuals. Columns 4&5 assess the capability of these methods on multimodal in-context comprehension tasks and advanced image-editing tasks (*e.g.,* multi-turn editing with consistent image fidelity). The final column clarifies whether the methods detach the comprehension and generation losses. We can see that all the existing MLLMs require the equivalence between pre- and post-MLLM visual-tokens, which causes the conflict between comprehension and generation training objectives. Thus, none of them can achieve synergy on complicated comprehension and generation tasks. In contrast, we propose morph-tokens to detach the textual and image reconstruction losses, where the pre-MLLM visual-tokens are not necessarily equal to the post-MLLM ones (Column 2&3), thus effectively resolving the conflict and achieving the synergy (Column 4&5). More detailed comparisons are provided in Section 4.

3. Method

We introduce the proposed morph-token-based MLLM in Figure 4. The detailed implementations are in Appendix B.

3.1. Encoder

As illustrated in Figure 4(a), given visual-tokens $\mathcal{V}$ extracted from an image, *i.e.*, by CLIP-ViT (Fang et al., 2023), our encoder is proposed to abstract these visuals by transforming them into morph-tokens $\mathcal{M}$, which serve as visual prompts for comprehension tasks. As shown in Figure 4(a), we use Q-former (Li et al., 2023b) to abstract $\mathcal{V}$ into token embeddings, which are quantized into discrete morph-tokens $\mathcal{M}$:

$$\mathcal{M} = \texttt{Quantizer}(\ \texttt{Qformer}(Q=\mathcal{Q}, K=\mathcal{V}, V=\mathcal{V})\ ) \tag{1}$$

where the arguments (Q, K, V) denote the query, key, and value of Q-former, $\mathcal{Q}$ is a set of query embeddings obtained from below.

As the CLIP-ViT visual tokens are merely flattened 2D patch features, the attention across such spatial visual-tokens is usually spuriously correlated (Wang et al., 2020). To remove such spatial confounding effect, inspired by (Wang et al., 2020), we integrate a deconfounder dictionary $\mathcal{D}$ to initialize the above query embeddings $\mathcal{Q}$, making the resultant morph-tokens behave more like natural language that is causal (Pearl & Mackenzie, 2018). Specifically, given a set of learnable query vectors $\mathcal{G}$, we initialize $\mathcal{D}$ as a learned dictionary of pre-trained ViT-VQGAN and adopt a single-layer Q-former to obtain $\mathcal{Q}$:

$$\mathcal{Q} = \texttt{Single-Qformer}(Q=\mathcal{G}, K=\mathcal{D}, V=\mathcal{D}). \tag{2}$$

3.2. Morph-token-based MLLM

After transforming $\mathcal{V}$ into $\mathcal{M}$, the morph-token-based MLLM, where the pre-MLLM morph-tokens $\mathcal{M}$ are not necessarily equal to the post-MLLM ones $\hat{\mathcal{M}}$. Specifically, as shown in Figure 4 (b&c), for comprehension tasks, $\mathcal{M}$ serves as visual prompts to instruct the MLLM to generate text-tokens $\mathcal{Y}$; for generation tasks, MLLM produces another set of post-MLLM morph-tokens $\hat{\mathcal{M}}$ which recover the visual features lost by $\mathcal{M}$, thus $\hat{\mathcal{M}}$ can generate images. In this way, $\mathcal{M}$ and $\hat{\mathcal{M}}$ effectively resolve the conflicting objectives of visual comprehension and generation. However, $\hat{\mathcal{M}}$ *per se* cannot yet generate images of high fidelity because it is not reasonable to force the MLLM to recover all the high-frequency visual details. Therefore, we need to further decode $\hat{\mathcal{M}}$ into lower-level visual-tokens $\mathcal{X}$ that can be finally decoded back to pixels by VQGAN. This decoder is introduced below.

3.3. Decoder

Our design philosophy is to allow the lower-level visual tokens to auto-regressively generate their own visual distributions. This design aims to disentangle the different distributions of natural language and visual tokens (Figure 2(c)). Such disentanglement further helps in resolving the conflict during MLLM training. As illustrated in Figure 4(c),

$$\text{Image} = \texttt{VQGAN}(\mathcal{X}),\ \ \mathcal{X} = \texttt{Decoder}(\hat{\mathcal{M}}), \tag{3}$$

where $\hat{\mathcal{M}}$ serves as a higher-level visual prompt to instruct `Decoder`, a decoder-only Transformer that decodes $\mathcal{X}$ which can be fed into a pre-trained VQGAN decoder to generate an image.

3.4. Training Strategy

We detail the 3-stage strategy, as illustrated in Figure 3, for training the morph-token-based MLLM.

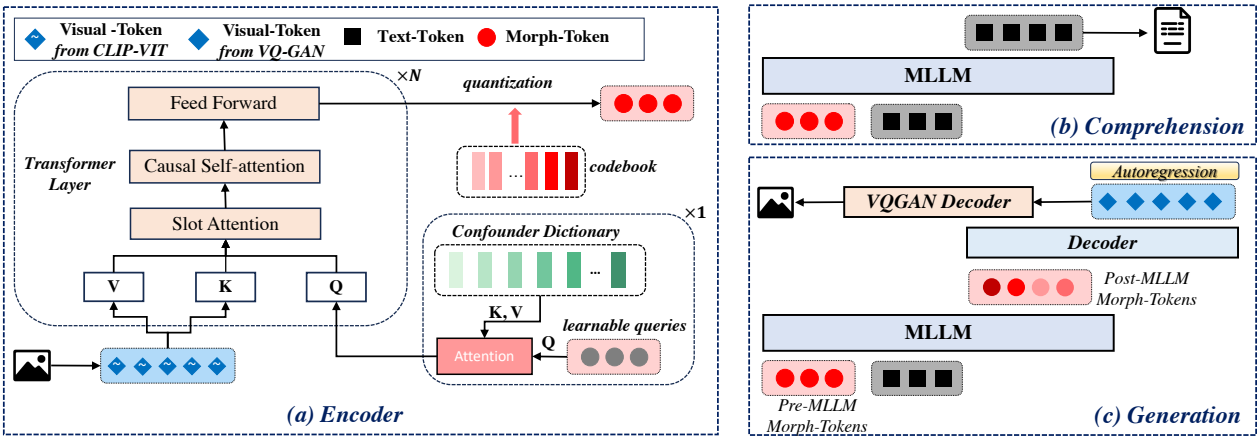

Figure 4. (a) The encoder introduced in Section 3.1. (b) For comprehension tasks, pre-MLLM morph-tokens instruct MLLM to generate texts. (c) For generation tasks, post-MLLM morph-tokens are first decoded into a lower-level visual-tokens (blue), introduced in Section 3.2. Then, they are fed into the VQ-GAN decoder to generate high-fidelity images.

Stage 1: Initialization. We aim to extend the token vocabulary of a pre-trained Vicuna (Chiang et al., 2023), transitioning it from LLM to MLLM. We use ∼30M image-text pairs, which are concatenated in two formats, *i.e.*, $\langle \mathcal{M}, \mathcal{Y} \rangle$ and $\langle \mathcal{Y}, \mathcal{M} \rangle$. As shown in Figure 1(a), we fine-tune the morph-token encoder and the LLM to maximize the auto-regressive likelihood of the two-format concatenated tokens:

$$\underset{\theta_{Enc},\, \theta_{LLM}}{\arg\max} \; \log P\left(t_i \in \{\mathcal{M}, \mathcal{Y}\} \mid t_{<i} \in \{\mathcal{M}, \mathcal{Y}\}\right) \tag{4}$$

where t_i denotes a morph-/text-token, θ_{Enc} is the encoder parameters, and θ_{LLM} denotes the LoRA (Hu et al., 2021) parameters attached to Vicuna. In particular, we set the token length: $|\mathcal{M}| = 32$ and $|\mathcal{Y}| = 512$. Recall that although this stage requires equality between pre- and post-MLLM morph-tokens, there is no conflict due to the absence of a visual generation objective. The resultant vocabulary size of our MLLM is 8,192 morph-tokens and 32,000 text-tokens.

Stage 2: Auto-encoding Morph-Tokens. We use the same image-text pairs. As shown in Figure 3(b), for image-captioning comprehension task, we use the token format $\langle \mathcal{M}, \mathcal{Y} \rangle$, where $\mathcal{M}$ serves as a visual prompt to instruct the MLLM to generate $\mathcal{Y}$ auto-regressively:

$$\underset{\theta_{Enc},\, \theta_{LLM}}{\arg\max} \; \log P\left(t_i \in \mathcal{Y} \mid t_{<i} \in \{\mathcal{M}, \mathcal{Y}\}\right). \tag{5}$$

For text-to-image generation task, we use the token format $\langle \mathcal{Y}, \mathcal{M} \rangle$ to feed into MLLM that generate $\hat{\mathcal{M}}$ auto-regressively. Then, $\hat{\mathcal{M}}$ is decoded into $\mathcal{X}$ by Eq. (3) auto-regressively:

$$\underset{\theta_{Enc},\, \theta_{LLM},\, \theta_{Dec}}{\arg\max} \; \log P\left(x_i \in \mathcal{X} \mid x_{<i} \in \mathcal{X}, \{\hat{m}\}_{j=1}^{N} \in \hat{\mathcal{M}}\right) \tag{6}$$

where θ_{Dec} denotes the parameters of the decoder. Recall that the above reconstruction loss does not impose $\hat{\mathcal{M}} = \mathcal{M}$. Thus, the training objectives of Eq. (5) and Eq. (6) do not conflict. During inference, if the generated $|\hat{\mathcal{M}}| < |\mathcal{M}|$, we complete it with special $\langle Emp \rangle$ tokens; if $|\hat{\mathcal{M}}| > |\mathcal{M}|$, we trim it to $|\mathcal{M}|$. This stage can be viewed as an auto-encoding process: Image→$\mathcal{V}$→$\mathcal{M}$→$\hat{\mathcal{M}}$→$\mathcal{X}$→Image, unique in that it does not have a fixed morph-token bottleneck.

Stage 3: Instruction Tuning. Besides the image-text pairs used in the above two stages, we use extensive interleaved image-text data as "`<Instructions>`" to enhance the MLLM's comprehension and generation capabilities in complex scenarios, *e.g.*, `<Instructions>`$= \langle \mathcal{M}, \mathcal{Y}, \mathcal{M} \rangle$. We use the instruction format "`USER: <Instructions> ASSISTANT: <Answers>.`", where "`USER:`" and "`ASSISTANT:`" are prompt tokens, and the training loss is as the same as Stage 2. See Appendix C for the instruction examples of diverse tasks.

4. Experiments

Through instruction-tuning on extensive supervised image-text data from diverse tasks (e.g., text-to-image generation, image editing, image QA, multi-image understanding), our model evolves into a versatile multimodal generalist, excelling in zero-shot vision-language comprehension and synthesis tasks. We conduct thorough experiments across a wide range of vision-language tasks, making comparisons primarily with MLLMs (Koh et al., 2023; Sun et al., 2023b; Ge et al., 2023; Jin et al., 2023) designed for both visual comprehension and generation. And we also compare with some widely-used MLLMs (Alayrac et al., 2022; Li et al., 2023b; Dai et al., 2023; Zhu et al., 2023; Liu et al., 2023b; Ye et al., 2023) that specialize solely in comprehension tasks. For detailed experimental setups and implementation specifics, please refer to Appendix C.

Models	Size	Image Gen	Image caption		Image QA				Video QA		MME Bench	
			NoCaps	Flickr30K	GQA	VSR	ICONQA	HM	MSVDQA	MSRVTTQA	Perception	Cognition
Flamingo	9B	✗	-	61.5	-	-	-	57.0	30.2	13.7	-	-
BLIP-2	11B	✗	98.4	73.7	44.6	68.2	45.4	52.0	34.4	17.4	1293.8	290.0
InstructBlip	11B	✗	120.0	83.5	47.9	65.6	**51.2**	54.1	44.3	25.6	1212.8	291.8
MiniGPT4	13B	✗	-	-	30.8	41.6	37.6	-	-	-	581.7	144.3
LLaVA	13B	✗	-	-	41.3	51.2	43.0	-	-	-	**1531.3**	295.4
mPlug-Owl	13B	✗	-	85.1	56.1	-	-	-	42.4	23.6	1450.2	313.2
Emu-I	13B	-	106.8	80.9	46.0	53.9	42.9	56.2	37.0	21.2	660.9	257.9
Emu2-Chat	37B	-	119.8	86.0	**65.1**	50.4	49.9	57.2	49.0	31.4	1315.7	307.5
Seed-LLaMA	8B	✓	90.4	66.7	34.8	45.2	36.0	50.4	45.2	35.3	736.7	235.5
LaVIT	7B	✓	114.2	83.0	46.8	60.4	36.8	53.0	-	-	997.9	240.7
Ours	8B	✓	**124.0**	**87.5**	56.8	**69.8**	47.6	**62.0**	**50.9**	**37.2**	1477.7	**389.3**

Table 2. Comparison for multimodal comprehension. "Image Gen" denotes whether the model can generate images besides texts (Emu-I and Emu2-Chat only possess the image generation capability in the versions prior to instruction-tuning).

Models	MMD	VST	VRI	MMC	KGQA	TRQA	MMR
Flagmingo	16.9	24.2	13.9	21.7	32.0	30.6	41.6
Blip-2	26.1	21.3	10.7	17.9	39.2	33.5	39.7
InstructBlip	33.6	24.4	11.5	21.2	47.4	44.4	48.6
MiniGPT-4	13.7	17.1	8.0	16.6	30.3	26.4	43.5
LLaVA	7.8	10.7	8.3	15.9	36.2	28.3	41.5
mPlug-Owl	12.7	19.3	5.4	16.3	33.3	32.5	42.5
VPG-C	**37.5**	25.2	25.9	22.2	48.6	44.9	50.3
Emu-I	25.6	16.1	13.4	23.1	46.4	32.2	42.6
Emu2-Chat	26.8	19.8	13.6	19.3	54.6	44.2	46.7
Seed-LLaMA	10.4	15.7	11.5	18.5	30.9	33.3	44.6
LaVIT	36.5	25.5	10.8	26.7	38.0	38.2	45.8
Ours	32.2	**27.4**	**27.4**	**28.0**	**56.4**	**47.7**	**54.9**

Table 3. Average results of zero-shot evaluation on each task category of DEMON Benchmark.

4.1. Zero-shot Multimodal Comprehension

Image Caption and VQA. We first evaluate our model on a wide range of academic benchmarks including image captioning and image/video question answering datasets. As shown in Table 2, our model achieves competitive performance in both image and video understanding tasks. Specifically, (1) Compared to specialized visual comprehension MLLMs like InstructBlip (Dai et al., 2023), as well as models like Emu2-Chat that primarily concentrate on comprehension tasks with significantly larger parameter scales, our method simultaneously maintains the image generation capabilities and achieves enhanced multimodal comprehension capabilities. For instance, our model consistently attains higher Cider scores in image captioning tasks and performance in Video QA tasks, which requires the understanding of multiple images. (2) Furthermore, Compared to previous SOTA models (Seed-LLaMA and LaVIT) which are capable of both visual generation and understanding, our model consistently exhibits stronger performance in all these image captioning and Visual Question Answering benchmarks, which underscores the robust multimodal comprehension capacity of our model.

MLLM-oriented Comprehension Benchmarks. Our zero-shot evaluation also encompasses recent MLLM-oriented comprehension benchmarks, including MME (Fu et al., 2023a) and DEMON benchmark (Li et al., 2023c). We have the following observations: (1) The results on MME in Table 2 underscore the strong generalizability of our model to follow a diverse range of single-image instructions. Especially when compared to similar MLLMs with certain image generation capabilities, our model demonstrates a significant advantage in both perception and cognition abilities. (2) Table 3 showcases the superior performance of our model on the DEMON benchmark, which is specifically designed to evaluate a model's capability of in-context learning on following demonstrative instructions. And our model outperforms the previous SOTA model in the DEMON benchmark, i.e., VPG-C (Li et al., 2023c), across the majority of task categories. For instance, we achieve performance improvements of 5.6% in multi-modal cloze (MMC) tasks and 7.8% in knowledge-grounded image QA (KGQA) tasks compared to VPG-C, which underscores our advanced ability to associate interleaved text-image inputs for stronger in-context understanding.

4.2. Zero-shot Image Synthesis

Text-to-Image Generation. To evaluate our model's capabilities in zero-shot image synthesis, we first evaluate the text-to-image generation on MS-COCO(Lin et al., 2014) (30K randomly sampled data from validation set, and 5K data in karpathy test set) and Flickr30K (Young et al., 2014) (1K data in test set), and compute pair-wise CLIP similarity score as the evaluation metric following previous works (Koh et al., 2023; Ge et al., 2023). As shown in Table 4, compared to all other existing MLLMs, the images generated from textual descriptions by our method consistently exhibit higher similarity with the ground-truth images. This highlights that our model facilitates better vision-text alignment, thereby effectively transforming text prompts into more relevant images.

Image Editing. We further evaluate our model in zero-shot instruction-based image editing across three datasets: EVR (Tan et al., 2019), MA5k (Shi et al., 2021), and Mag-

Models	COCO *Karpathy test*	COCO *val30k*	Flickr *test*
GILL	68.4	67.5	65.2
Emu	65.6	66.5	64.8
Emu2-Gen	68.6	67.6	64.9
SEED-LLaMA	68.2	70.7	65.6
LaVIT	-	68.4	63.5
Ours	**70.6**	**72.2**	**68.8**

Table 4. Zero-shot Evaluation of text-to-image generation.

	EVR		MA5K		MagicBrush	
Method	L1↓	CVS↑	L1↓	LPIPS↓	L1↓	CVS↑
InsPix2Pix	18.9	81.4	17.6	35.9	10.1	85.2
LGIE	15.9	82.0	14.4	32.7	8.4	88.9
MGIE	16.3	81.7	**13.3**	29.8	8.2	**91.1**
Gill	31.8	65.0	27.4	44.3	28.3	75.2
Emu	30.7	69.2	27.2	43.2	27.9	78.5
Emu2-gen	22.8	80.3	20.5	28.4	19.9	85.7
SEED-LLaMA	28.4	72.3	24.6	39.0	24.5	80.9
LaVIT	26.8	73.8	25.1	36.9	25.3	81.1
Ours	**15.3**	**82.6**	14.6	**27.9**	**7.6**	87.9

Table 5. Zero-shot image editing results. The first three rows consist of models specialized in image editing. The best results among MLLMs capable of both multimodal comprehension and generation are underline.

icBrush (Zhang et al., 2023). Following Fu et al. (2023b), for EVR and MagicBrush, we treat the standard pixel difference (L1) and visual feature similarity from the CLIP visual encoder (CVS) between generated images and ground-truth goals as the evaluation metrics. For MA5K, we utilize L1 and Learned Perceptual Image Patch Similarity (LPIPS Zhang et al., 2018) as the evaluation metrics. The experimental results are shown in Table 5, leading to the following observations:

First, Across all datasets and metrics in image editing, our model significantly outperforms existing Multimodal MLLMs that possess unified multimodal comprehension and generation capabilities. Moreover, when compared to image-editing specialists (e.g., InsPix2Pix Brooks et al., 2023, MGIE Fu et al., 2023b), our model still achieves stronger performance. It consistently surpasses InsPix2Pix and shows greater efficacy than previous SOTA specialists, i.e., MGIE across most datasets. For instance, on the EVR dataset, our approach achieves lower L1 scores and higher CLIP similarity compared to MGIE. **Second**, compared with other MLLMs such as Emu2-Gen, our model demonstrates a more pronounced superiority in terms of the L1 score than it does on the CVS metric. We find that this disparity is attributed to the inability of existing MLLMs to maintain image fidelity throughout the editing process. Therefore, while models like Emu2 still achieve decent CVS scores, the performance in terms of L1 scores is not ideal. In contrast, our model not only effectively comprehends instructions to execute image editing, but also well preserves image fidelity, which leads to a significant advantage in L1 scores over other MLLMs.

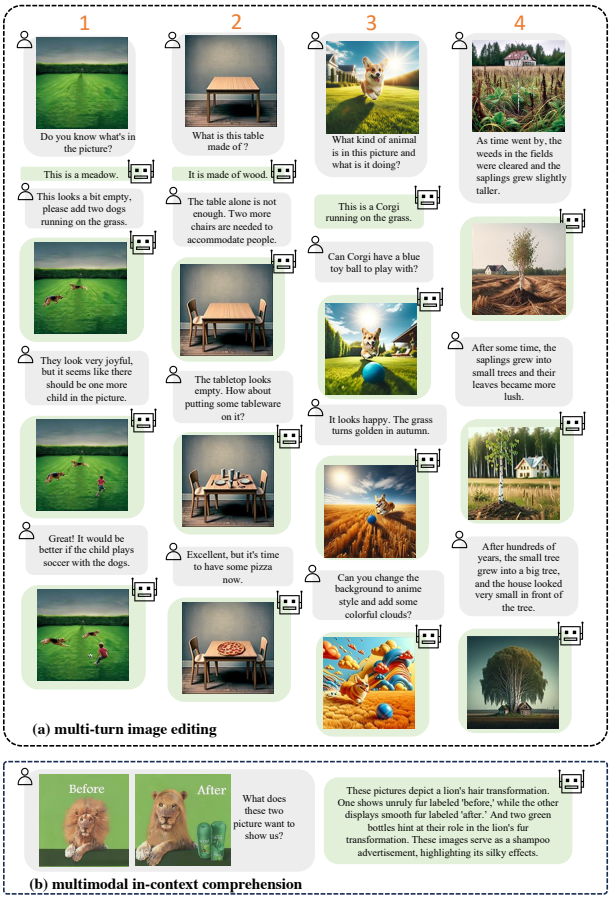

Figure 5. Qualitative results on multi-turn image editing and multimodal in-context comprehension.

4.3. Emergent Abilities

Multi-turn Image Editing. As shown in Figure 5.a, in multi-turn image editing scenarios, our model is capable of effectively interpreting diverse user instructions to edit the image, while ensuring the preservation of image fidelity. Visual objects not intended for alteration retain a high level of consistency before and after editing.

Mulmodal In-context Learning. Our model also exhibits an advanced capability for multimodal in-context learning. In comprehension tasks (Figure 5.b), it can keenly identify the connections between input images and provides insightful analyses. Moreover, in generation tasks (Figure 6), it can effectively understand the complete meaning of interleaved image-text inputs and engages in compositional image generation following instructions. Even in the absence of any natural language instructions, merely provided with several image pairs, our model is capable of inferring the patterns of change between images to understand the task requirements, thereby generating the desired image.

Figure 6. Qualitative results on multimodal in-context generation.

4.4. In-Depth Analysis

Impact of "Morph" . To explore the impact of morph tokens, we eliminate the design of "morph", and train the following ablation models which require the equivalence between pre- and post-MLLM visual tokens: **(1) detail-detail**: following TEAL (Yang et al., 2023), both pre- and post-MLLM visual tokens contain complete visuals. **(2 & 3) abstr-abstr (SD & VQ)**: both pre- and post-MLLM visual tokens contain abstracted semantics. And image generation is facilitated either through the use of SD models (following SEED-LLaMA) or via another decoder-only transformer to reconstruct the detailed visuals that can be converted into the image with a VQ-GAN decoder. The implementation details are shown in Appendix C. We evaluate their performance on Demon benchmark and image-editing tasks (with L1 as the evaluation metric), as shown in Table 7 and 8.

Firstly, the results of Row1 show that when pre-MLLM tokens contain detailed visual semantics, MLLMs not only fail to facilitate effective visual comprehension (as evidenced by unsatisfactory performance on Demon), but also fall short in image editing, which necessitates a comprehensive understanding of the image being edited. **Secondly**, The results of Row2 and Row3 show that when both pre- and post-MLLM visual tokens embody abstract semantics, MLLMs firstly struggle to effectively conduct image editing tasks without the support of visually-complete post-MLLM tokens. Moreover, forcing them to possess visual generation capabilities also notably compromises their effectiveness in executing complex visual comprehension tasks like DEMON.

Model	Image → Text			Text → Image			
	R@1	R@5	R@10	R@1	R@5	R@10	R@m
BLIP-2	81.9	98.4	99.7	82.4	96.5	98.4	92.9
SEED	**91.0**	99.5	**100.0**	79.3	94.8	97.1	93.6
LaVIT	83.0	99.2	99.7	78.3	96.2	97.5	92.3
Ours	88.8	**99.7**	**100.0**	**85.2**	**97.1**	**98.7**	**95.0**

Table 6. Evaluation of Image-Text Retrieval on Flickr30K.

Effectiveness of Individual Components. To further investigate the effectiveness of individual components, we train the following ablation models and also evaluate on Demon benchmark and image-editing tasks: **(1) w/o decoder**: we force the MLLM to recover all the high-frequency visual details, where post-MLLM visual tokens are directly fed into a VQ-GAN decoder for image generating. The results of Row4 in Table 7 and 8 show that it is quite difficult for MLLM to directly autoregress such lower-level visual-tokens that can be finally decoded back to pixels by VQ-GAN. And an additional decoder is essential to alleviate its burden in recovering the lost visual features. **(2) w/o deconfounded**: we remove the deconfounding design in the encoder. The results of Row 5 in Table 7 and 8 demonstrate

	Models	MMD	VST	VRI	MMC	KGQA	TRQA	MMR
Morph-token		**32.2**	**27.4**	**27.4**	**28.0**	**56.4**	**47.7**	**54.9**
1	detail-detail	7.3	8.6	8.2	16.6	30.2	24.8	34.8
2	abstr-abstr (SD)	25.8	17.3	11.3	22.7	38.8	33.5	45.0
3	abstr-abstr (VQ)	25.2	20.4	13.2	19.8	37.9	32.7	46.1
4	w/o decoder	20.9	16.4	12.8	18.0	35.6	30.4	41.3
5	w/o deconfound	31.8	25.7	25.3	26.2	54.7	44.8	50.9
6	continuous	31.7	26.3	26.1	27.0	54.3	45.2	52.5

Table 7. Ablation results on DEMON Benchmark.

	Models	EVR↓	MA5K↓	MagicBrush↓
Morph-token		**15.3**	**14.6**	**7.6**
1	detail-detail	34.1	31.2	32.5
2	abstr-abstr (SD)	27.7	26.6	27.8
3	abstr-abstr (VQ)	25.2	24.4	23.9
4	w/o decoder	31.7	28.9	30.3
5	w/o deconfound	18.5	18.2	12.2
6	continuous	16.1	15.8	9.7

Table 8. Ablation results on image editting with L1 score as the evalution metric.

that deconfounding enables visual tokens to behave more like natural language, which leads to enhanced performance in both visual comprehension and generation.

Imapct of Discrete Morph-Tokens. We also validate the superiority of quantizing visual abstraction into discrete tokens as the pre-MLLM morph-tokens. We eliminate the process of visual quantization and also change the optimization objective of visual tokens as regressing the next token with MSE loss in stage 1. As indicated in Row 6 of Table 7 and 8, it results in a degradation of model performance for both comprehension and generation tasks. We argue a key reason for this is that discretization aligns visual tokens more closely with the attributes of natural language. Then in training stage 1, we can employ a uniform objective (cross-entropy loss) for both vision and language, which better improves their alignment within the LLM (Jin et al., 2023), thereby facilitating a more effective transition of the LLM into an MLLM.

Image-Text Retrieval with our Encoder. To further validate that our proposed encoder can effectively translate the non-linguistic image into a sequence of morph-tokens that behave more like natural language, we further evaluate the performance of our encoder in zero-shot image-text retrieval tasks, utilizing the Flickr30K dataset with Recall@K (R@K) as the evaluation metric. We compare with the visual encoders in SEED-LLaMA (Ge et al., 2023) and LAVIT (Jin et al., 2023), as well as the Qformer in BLIP2 (Li et al., 2023b). As shown in Table 6, our encoder surpasses those from existing MLLMs across the majority of metrics in text-image retrieval tasks, while also achieving superior average performance (R@m). This validates that our encoder can better abstract visual semantics to align with text representations, thereby effectively alleviating the modality gap.

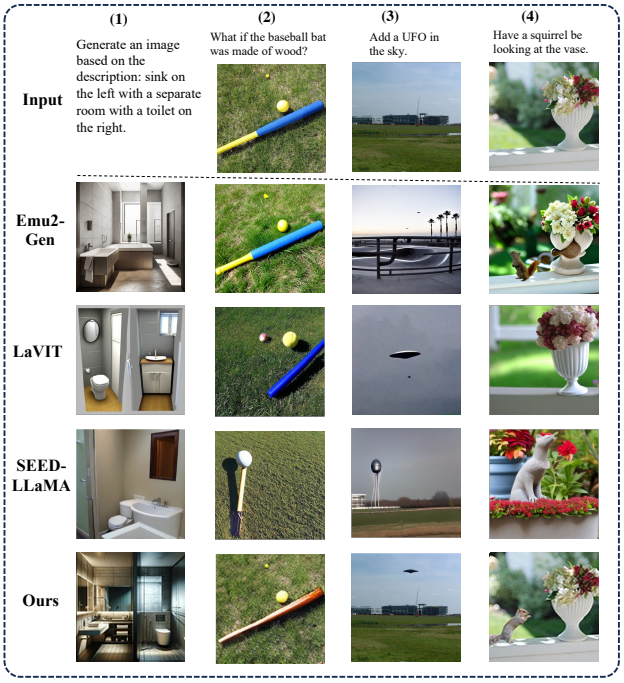

Figure 7. Qualitative comparison on image synthesis.

Qualitative Comparison on Image Synthesis. Figure 7 presents a qualitative comparison between our model and other MLLMs in the context of image generation tasks. We can see that in the text-to-image generation task when the textual descriptions involve spatial relationships, images generated by MLLM often inaccurately represent the relationships among specific visual objects. In contrast, our model precisely generates the image that adheres to the predetermined specifications detailed in the prompt, reflecting the intended spatial relationships. While in image editing scenarios, it can be observed that our approach well understands the instructions, while also effectively preserving image fidelity, which is rarely observed in prior works.

5. Conclusion

In this paper, we propose Morph-Tokens to resolve the conflicting training objectives between visual comprehension and generation– "morph" implies a transformation where the pre-MLLM visual tokens are not necessarily equal to the post-MLLM ones. The pre-MLLM tokens are abstract semantics, serving as visual prompts for comprehension tasks while the post-MLLM tokens are visually complete tokens for image generation. We further propose a 3-stage training strategy, detaching the textual and image reconstruction losses with our morph-tokens. After training, our model showcases notable zero-shot performance on a broad range of comprehension and generation tasks, also exhibiting extensive emergent abilities such as consistently preserving image fidelity in image editing scenarios.

Acknowledgements

This work was supported by the Key Research and Development Projects in Zhejiang Province (No. 2024C01106), the NSFC (No. 62272411), the National Key Research and Development Project of China (2018AAA0101900), the Tencent WeChat Rhino-Bird Special Research Program (Tencent WXG-FR-2023-10), and Research funding from FinVolution Group.

Impact Statement

Ethical Impacts. This study does not raise any ethical concerns. The research does not involve subjective assessments or the use of private data. Only publicly available datasets are utilized for experimentation.

Expected Societal Implications. A major societal concern with this technology lies in its potential for misuse, particularly in fabricating unauthorized images that could lead to misinformation, privacy breaches, and other damaging consequences. To counter these threats, it is crucial to develop strong ethical standards and implement ongoing surveillance.

The issue highlighted is not unique to our method but is prevalent across different techniques for multi-concept customization. A practical approach to mitigating these risks could involve the use of adversarial technologies designed to reverse unauthorized modifications. Additionally, embedding imperceptible watermarks in created images could act as a preventative measure against abuse and guarantee that their use is properly attributed.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. nocaps: novel object captioning at scale. In *ICCV*, pp. 8948–8957, 2019.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Cao, M., Li, S., Li, J., Nie, L., and Zhang, M. Image-text retrieval: A survey on recent research and development, 2022.

Chen, D., Pan, K., Wang, G., Zhuang, Y., and Tang, S. Improving vision anomaly detection with the guidance of language modality. *arXiv preprint arXiv:2310.02821*, 2023.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality, 2023. URL https://vicuna.lmsys.org.

Christoph, S., Andreas, K., Richard, V., Theo, C., and Romain, B. Laion coco: 600m synthetic captions from laion2b-en. [EB/OL], 2022. https://laion.ai/blog/laion-coco/.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835, 2021.

Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023a.

Fu, T.-J., Hu, W., Du, X., Wang, W. Y., Yang, Y., and Gan, Z. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023b.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Ge, Y., Ge, Y., Zeng, Z., Wang, X., and Shan, Y. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

Ge, Z., Huang, H., Zhou, M., Li, J., Wang, G., Tang, S., and Zhuang, Y. Worldgpt: Empowering llm as multimodal world model. *arXiv preprint arXiv:2404.18202*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Jin, Y., Xu, K., Chen, L., Liao, C., Tan, J., Chen, B., Lei, C., Liu, A., Song, C., Lei, X., et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.

Koh, J. Y., Fried, D., and Salakhutdinov, R. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023a.

Li, J., He, X., Wei, L., Qian, L., Zhu, L., Xie, L., Zhuang, Y., Tian, Q., and Tang, S. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022a.

Li, J., Xie, J., Qian, L., Zhu, L., Tang, S., Wu, F., Yang, Y., Zhuang, Y., and Wang, X. E. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3032–3041, 2022b.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

Li, J., Pan, K., Ge, Z., Gao, M., Zhang, H., Ji, W., Zhang, W., Chua, T.-S., Tang, S., and Zhuang, Y. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *arXiv preprint arXiv:2308.04152*, 2023c.

Li, W., Xu, X., Xiao, X., Liu, J., Yang, H., Li, G., Wang, Z., Feng, Z., She, Q., Lyu, Y., et al. Upainting: Unified text-to-image diffusion generation with cross-modal guidance. *arXiv preprint arXiv:2210.16031*, 2022c.

Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., and Jia, J. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, F., Emerson, G. E. T., and Collier, N. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., and Zhu, S.-C. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS Track on Datasets and Benchmarks*, 2021.

Pan, J., Sun, K., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al. Journeydb: A benchmark for generative image understanding. *arXiv preprint arXiv:2307.00716*, 2023a.

Pan, K., Li, J., Song, H., Lin, J., Liu, X., and Tang, S. Self-supervised meta-prompt learning with meta-gradient regularization for few-shot generalization. *arXiv preprint arXiv:2303.12314*, 2023b.

Pan, K., Li, J., Wang, W., Fei, H., Song, H., Ji, W., Lin, J., Liu, X., Chua, T.-S., and Tang, S. I3: Intent-introspective retrieval conditioned on instructions, 2024.

Pearl, J. and Mackenzie, D. *The book of why: the new science of cause and effect*. Basic books, 2018.

Qian, L., Li, J., Wu, Y., Ye, Y., Fei, H., Chua, T.-S., Zhuang, Y., and Tang, S. Momentor: Advancing video large language model with fine-grained temporal reasoning, 2024.

Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., and Hoque, E. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, pp. 2359. NIH Public Access, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Shi, J., Xu, N., Xu, Y., Bui, T., Dernoncourt, F., and Xu, C. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13590–13599, 2021.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. Generative multimodal models are in-context learners. 2023a.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023b.

Tan, H., Dernoncourt, F., Lin, Z., Bui, T., and Bansal, M. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019.

Wang, T., Huang, J., Zhang, H., and Sun, Q. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10760–10770, 2020.

Wang, T., Lin, K., Li, L., Lin, C.-C., Yang, Z., Zhang, H., Liu, Z., and Wang, L. Equivariant similarity for vision-language foundation models. *arXiv preprint arXiv:2303.14465*, 2023.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022a.

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022b.

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1645–1653, 2017.

Yang, Z., Zhang, Y., Meng, F., and Zhou, J. Teal: Tokenize and embed all for multi-modal large language models. *arXiv preprint arXiv:2311.04589*, 2023.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Yu, L., Lezama, J., Gundavarapu, N. B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A. G., et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023a.

Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023b.

Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Zhang, K., Mo, L., Chen, W., Sun, H., and Su, Y. Magicbrush: A manually annotated dataset for instruction-guided image editing. *arXiv preprint arXiv:2306.10012*, 2023.

Zhang, Q., Zhang, Y., Wang, H., and Zhao, J. Recost: External knowledge guided data-efficient instruction tuning. *arXiv preprint arXiv:2402.17355*, 2024a.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhang, X., Li, J., Chu, W., Hai, J., Xu, R., Yang, Y., Guan, S., Xu, J., and Cui, P. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024b.

Zhao, B., Wu, B., and Huang, T. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. Detailed Comparison with Existing MLLMs.

MLLMs (Li et al., 2023b; 2022a; Zhu et al., 2023; Qian et al., 2024) aim to serve as a general-purpose assistant to perform various vision-language tasks (Li et al., 2022b; Chen et al., 2023) with strong generalization ability (Zhang et al., 2024b; Pan et al., 2023b). To drive both multimodal comprehension and generation tasks in a unified, token-based, auto-regressive framework, existing approaches (Sun et al., 2023b; Yang et al., 2023; Ge et al., 2023; 2024; Jin et al., 2023) typically leverage a tokenizer to encode images into visual tokens and feed them into an MLLM for vision-language comprehension. In terms of visual generation, the post-MLLM visual tokens are further used to generate target images, which is achieved either through a pre-trained VQVAE decoder dedicated to image pixel reconstruction, or via Stable Diffusion (SD (Rombach et al., 2021)) models.

Specifically, models like TEAL (Yang et al., 2023) and VideoPoet (Kondratyuk et al., 2023), utilize the VQ-VAE encoder as the tokenizer (e.g., VQGAN (Esser et al., 2021) and MAGVIT-v2(Yu et al., 2023a)), encoding images into visual tokens with detailed semantics for unified auto-regression within MLLMs. The post-MLLM visual tokens, which are visually complete, can then be directly converted into images using the corresponding VQ-VAE decoder. However, as these tokens preserve low-level visual details, while being suitable for visual generation, they substantially impede the capability of visual comprehension.

Another line of work (Ge et al., 2023; Sun et al., 2023b) attempts to first extract abstracted visuals for comprehension, where visual tokens and text tokens undergo a unified autoregression within the MLLM. And then akin to "Textual Inversion" (Gal et al., 2022), the post-MLLM visual tokens are further aligned into the condition embedding space of existing SD model (e.g., through MSE loss), facilitating SD models to generate the image. However, two significant challenges arise: (1) post-MLLM visual tokens, similar to pre-MLLM ones, also encapsulate abstract semantics that are insufficient for image generation. To address the conflicting objectives, methods like Emu2 (Sun et al., 2023a) opt to train separate models for distinct purposes: Emu-gen for generation and Emu-chat for comprehension. (2) Moreover, existing SD models mainly focus on simple scene image generation, trained with coarse-grained conditions (Li et al., 2022c). For instance, the unCLIP-SD (Rombach et al., 2022) used in SEED-LLaMA (Ge et al., 2023) utilizes CLIP image embeddings as the condition, which contain only modality-shared information and often overlook the modality-specific knowledge derived from multimodal comprehension (Liang et al., 2022; Dong et al., 2023). Therefore, it is challenging for SD models to achieve detailed control during image generation which rarely preserves image fidelity, especially in image editing scenarios. And comparing the results of rows 2 and 3 in Table 8 further substantiates this point. We can see that introducing an additional decoder for image reconstruction (Row3 in Table 8) yields better image editing performance (measured by L1 score) compared to simply aligning with the modality-shared condition embedding of SD models (Row2 in Table 8), where the decoder allows the lower-level visual tokens to autoregressively generate their own visual distributions.

In contrast to these MLLMs, we propose morph-tokens to detach the textual and image reconstruction losses, where the pre-MLLM visual tokens (with abstract semantics) are not necessarily equal to the post-MLLM ones (with visually-complete semantics), effectively resolving the conflicting objectives between comprehension and generation. Moreover, we employ a deconfounded Qformer as the encoder, enabling pre-MLLM visual tokens to behave more like natural language compared with existing tokenizers (e.g., SEED(Ge et al., 2023)). And we further introduce another decoder to alleviate the burden of MLLM for visual semantic recovering, consequently fostering a synergy between visual comprehension and generation.

B. Detailed Implementations of our Framework

We mainly introduce the detailed implementations of the encoder. Given an image, it is first transformed into a sequence of visual tokens $\mathcal{V}$ via CLIP-ViT, with each token encapsulating patch-level visual details. And the role of the encoder is to abstract these visuals by transforming them into morph-tokens. To achieve this, we propose a novel deconfounded Qformer to implement our encoder, eliminating the spatially spurious correlation in vision, enabling the resultant morph-tokens to behave more like natural language.

Firstly, following Qformer, we introduce a set of learnable query vectors (here we refer to as group tokens), and employ an attention-based method for semantic aggregation. Specifically, in contrast to the vanilla Qformer, we implement two improvements: we upgrade the self-attention mechanism to causal self-attention, wherein each token exclusively attends to its preceding tokens, thus endowing the sequence with causal dependency. Furthermore, we replace the pivotal cross-attention computation in Qformer with slot-attention, which still utilizes the group tokens as the query and visual tokens as the key/value. Diverging from the traditional cross attention in transformer decoders, slot-attention performs normalisation

over queries, encouraging each visual token to be claimed by one of the group tokens, with the attention score $\mathbb{A}$ calculated as follows:

$$\mathbb{A} = \text{Softmax}_{qry}(f_{\mathcal{G}}(X)) = \text{Softmax}_{qry}(\frac{(\mathcal{G}W_q)(XW_k)^T}{\sqrt{scaled}}), \ \sum_j \mathbb{A}j, k = 1 \tag{7}$$

And then the output of the Qformer successfully encapsulates the desired visual abstraction. Post-processing through an additional MLP layer, the features of visual abstraction are then passed to a learnable codebook $\mathcal{C}$ and quantized into a sequence of discrete visual codes as the morph-tokens through nearest neighbors lookup.

Based on the above framework, we further introduce the design of deconfounding to enhance morph-tokens for emulating natural language. Unlike the sequential manner in which humans understand language, image comprehension typically involves capturing a holistic visual impression from several key areas, and then diverging into specific image details. Sequentially flattening 2D images into 1D features can result in spurious correlations between two spatial visual tokens, thereby confounding the semantic abstraction of specific visual objects. For instance, imagine an image where a boy is leisurely watching a disaster movie at home. Reading the image sequentially akin to text processing may lead to a confused understanding of virtual and real worlds, mistakenly placing the boy within the movie scene, consequently extracting incorrect information about him.

In order to screen out the existence of confounders and then eliminate their effect, we use a mental apparatus, **intervention**. As depicted in Figure 8, $\mathcal{V}$ corresponds to the visual tokens with detailed patch-level semantics, and $\mathcal{M}$ represents the abstracted visual token after semantic aggregation. $\mathcal{V} \rightarrow \mathcal{M}$ denotes the process of visual semantic abstraction. Additionally, the confounder $\mathcal{D}$ represents other image patches which also directly affect $\mathcal{M}$ during aggregation. Simultaneously, its existence may erroneously impact the semantic abstraction for $\mathcal{V}$, leading to spurious correlations by relying solely on the likelihood $P(\mathcal{M}|\mathcal{V})$, where the confounder introduces the observational bias via $P(d|\mathcal{V})$:

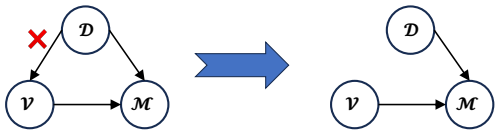

Figure 8. causal intervention.

$$P(\mathcal{M}|\mathcal{V}) := \sum_{d \in \mathcal{D}} P(\mathcal{M}|\mathcal{V}, d)\underline{P(d|\mathcal{V})} \tag{8}$$

To calculate the true causal effect between $\mathcal{V}$ and $\mathcal{M}$, we could intervene on $\mathcal{V}$ to cut off the causal link between $\mathcal{D}$ and $\mathcal{V}$, as shown in Figure 8(right). Utilizing Bayes rule on the revised graph, we have:

$$P(\mathcal{M}|do(\mathcal{V})) := \sum_{d \in \mathcal{D}} P(\mathcal{M}|\mathcal{V}, d)\underline{P(d)} \tag{9}$$

In contrast to Eq. (8), d is no longer influenced by $\mathcal{V}$. Consequently, the intervention compels $\mathcal{V}$ to fairly incorporate every d into the prediction of $\mathcal{M}$, in accordance with the prior probability $P(d)$.

To implement the causal intervention in Eq. (9), we should first include an additional confounder set $\mathcal{D}$ to enumerate the feature of different image patches, while it is impossible to attend to all image pathes in the real world. Fortunately, some works, *e.g.*, VIT-VQGAN (Yu et al., 2021) or DALLE (Ramesh et al., 2021), effectively quantize path-level embeddings from diverse images into a finite set of discretized latent codes within a learned codebook, providing a valuable resource to initialize the confounder dictionary.

As shown in Eq. (7), the critical step of visual semantic abstraction ($\mathcal{V} \rightarrow \mathcal{M}$) is facilitated through the slot attention, with a query-wise softmax is used to determine which clusters a low-level detailed visual token should be allocated to. Therefore, the implementation of causal intervention should be reflected in slot attention to upgrade the query-wise softmax, as delineated below:

$$P(\mathcal{M}|do(\mathcal{V})) := \mathbb{E}_{\mathbf{d}}[\text{Softmax}_{qry}(f_{\mathcal{G}}(\mathcal{V}, d))] \tag{10}$$

where $f(\cdot)$ calculates the logits of scaled dot-product attention. However, in this way $\mathbb{E}_{\mathbf{d}}$ requires expensive sampling with the cost of a network forward pass for all terms in $\mathcal{D}$. So we apply Normalized Weighted Geometric Mean (NWGM) to approximate the above expectation, effeciently moving the outer expectation into the Softmax operation as:

$$\mathbb{E}_{\mathbf{z}}[\text{Softmax}_{qry}(f_{\mathcal{G}}(\mathcal{V}, d))] \approx \text{Softmax}_{qry}(\mathbb{E}_{\mathbf{d}}[f_{\mathcal{G}}(\mathcal{V}, d)]) \tag{11}$$

To achieve the above approximation, we expand the query of slot attention as $\mathcal{Q} = \mathcal{G}W_q + \mathbb{E}_d[h_{\mathcal{G}}(d)]$, and then Eq. (11) can be derived as:

$$\begin{aligned} &\mathbb{E}_{\mathbf{z}}[\text{Softmax}_{qry}(f_{\mathcal{G}}(\mathcal{V}, d))] \\ &= \text{Softmax}_{qry}(\frac{(\mathcal{G}W_q + \mathbb{E}_{\mathbf{d}}[h_{\mathcal{G}}(d)])(\mathcal{V}W_k)^T}{\sqrt{scaled}}) \end{aligned} \tag{12}$$

Moreover, to compute $\mathbb{E}_{\mathbf{d}}[h_{\mathcal{G}}(d)]$, we also employ an attention-based mechanism, which, for convenience, we refer to as a single-layer Q-former (a module that includes only the computation of cross-attention). We treat the group tokens $\mathcal{G}$ as the query and the confounder dictionary $\mathcal{D}$ as both key and value. Through this, we derive an attention matrix $\mathcal{A}$ over each item in the dictionary. Then we can have $\mathbb{E}_{\mathbf{d}}[h_{\mathcal{G}}(d)] = \sum_z [A \odot \mathcal{D}]P(d)$, where $P(d)$ signifies the prior statistical probability and $\odot$ represents the element-wise product.

C. Experimental Details

C.1. Data

Pretraining Data. In stage 1 and stage 2, we select ∼30M image-text pairs from CC3M (Sharma et al., 2018) and Laion (Christoph et al., 2022), which are concatenated in two formats, *i.e.*, [text][image] and [image][text], facilitating the alignment between text and vision.

Instruction Tuning Data. During instruction tuning (Liu et al., 2023b; Zhang et al., 2024a), we incorporate a variety of tasks, outlined as follows: (1) Text-to-Image Generation: We employ datasets including JourneyDB (Pan et al., 2023a) and DiffusionDB (Wang et al., 2022b), utilizing a prompt template formatted as:"`USER: {caption} Generate an image based on the description. ASSISTANT: {image}`".

(2) Image editing: We employ datasets such as IPr2Pr (Brooks et al., 2023), utilizing a prompt template formatted as:"`USER: {image1} What will this image be like with the editing instruction: {instruction}. ASSISTANT: {image2}`".

(3) Image caption & Image QA & Video QA: We mainly leverage the held-in instruction-tuning datasets and corresponding instruction templates used in InstructBlip (Dai et al., 2023).

(4) Image Conversation: We employ the datasets including LLaVA (Liu et al., 2023b), SVIT (Zhao et al., 2023) with the prompt template formatted as: "`USER: {image} {question}. ASSISTANT: {answer}`".

(5) Multi-Image Understanding: We leverage GSD (Li et al., 2023a) as the training dataset, utilizing a prompt template formatted as:"`USER: This is the first image. {image1} This is the second image. {image2} {question } ASSISTANT: {answer}`".

Evaluation Data. For comprehension tasks, we first evaluate on a wide range of academic benchmarks, including NoCaps (Agrawal et al., 2019), Flickr30K (Young et al., 2014), GQA (Hudson & Manning, 2019), VSR (Liu et al., 2023a), ICONQA (Lu et al., 2021), HatefulMeme (Kiela et al., 2020), MSVDQA (Xu et al., 2017), and MSRVTTQA (Xu et al., 2017). The split of test sets and the evaluation metrics are aligned with those described in InstructBlip (Dai et al., 2023). Additionally, we also include some MLLM-oriented comprehension benchmarks, such as MME (Fu et al., 2023a) and the DEMON benchmark (Li et al., 2023c). For generation tasks, our evaluation encompasses both text-to-image generation and image editing. The former includes datasets of MS-COCO (Lin et al., 2014), (with 30K randomly sampled data from the validation set and 5K data from the Karpathy test set), and Flickr30K (Young et al., 2014) (with 1K data in the test set). For image editing, we evaluate the performance using datasets such as EVR (Tan et al., 2019), MA5k(Shi et al., 2021), and MagicBrush (Zhang et al., 2023). The partitioning of test sets and the evaluation metrics adhere to MGIE (Fu et al., 2023b).

C.2. Training.

We train the entire set of parameters for both the encoder and decoder. For the LLM, to enhance efficiency, we employ LoRA tuning (Hu et al., 2021) and together optimize the parameters of the decoder head layer due to the added visual words. With LoRA, we finetune W_q and W_v via low-rank adaptation. In our implementation, we set the rank, $r = 64$. utilize the AdamW optimizer coupled with a cosine learning rate scheduler. The hyperparameters for the AdamW optimizer are set with $\beta = (0.9, 0.999)$, and we apply a weight decay of 0.05. The training is conducted on 16xA800 GPUs. For the first two

stages, we train for 200,000 steps with a maximum learning rate as 1e-4. During instruction tuning, the model is trained for 100,000 steps with a maximum learning rate of 1e-5.

C.3. Ablation Model Implementation.

Here we give some implementation details of ablation models. (1) detail-detail: both pre- and post-MLLM visual tokens contain detailed semantics. Following TEAL, We integrate visual tokens from VQ-GAN encoder into the MLLM for unified auto-regression alongside text tokens, and directly convert the post-MLLM tokens into a specific image via VQ-GAN decoder. (2) abstr-abstr (SD): Both pre- and post-MLLM visual tokens contain abstract semantics. Utilizing our encoder to abstract the visuals, we leverage a unified auto-regressive objective for both textual and visual tokens within MLLM. Furthermore, following SEED-LLaMA (Ge et al., 2023), the post-MLLM visual tokens are aligned with the condition embedding of SD model (Rombach et al., 2022) (trained with MSE loss between token-embedding and the ground-truth condition embedding) (3) abstr-abstr (VQ): Based on the previous ablation model, we remove the SD model and instead train a decoder-only transformer, which auto-regressively predicts the complete visual token sequence that can be decoded into an image via VQGAN decoder (Esser et al., 2021), instructed by post-MLLM visual tokens.

Moreover, we also conduct image-text retrieval experiments (Cao et al., 2022; Pan et al., 2024). We adopt the dual-stream paradigm and incorporate the text encoder from BLIP2 (Li et al., 2023b). Concurrently, we learn a linear-projection layer using some image-text pairs from LAION, to align the output of the morph-encoder with that of the text encoder. We compare with the visual encoders in SEED-LLaMA (Ge et al., 2023) and LAVIT (Jin et al., 2023), as well as the Qformer in BLIP2 (we remove the image-text-matching re-rank module in BLIP2 to ensure a fair comparison).