VILA-U: A UNIFIED FOUNDATION MODEL INTEGRATING VISUAL UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

VILA-U is a **U**nified foundation model that integrates **V**ideo, **I**mage, **La**nguage understanding and generation. Traditional visual language models (VLMs) use separate modules for understanding and generating visual content, which can lead to misalignment and increased complexity. In contrast, VILA-U employs a single autoregressive next-token prediction framework for both tasks, eliminating the need for additional components like diffusion models. This approach not only simplifies the model but also achieves near state-of-the-art performance in visual language understanding and generation. The success of VILA-U is attributed to two main factors: the unified vision tower that aligns discrete visual tokens with textual inputs during pretraining, which enhances visual perception, and autoregressive image generation can achieve similar quality as diffusion models with high-quality dataset. This allows VILA-U to perform comparably to more complex models using a fully token-based autoregressive framework.

1 INTRODUCTION

In recent years, large language models (LLMs) have demonstrated superior capabilities in various language tasks. Their appealing properties like instruction following, zero-shot generalization, and few-shot in-context learning motivate researchers to combine them with vision models to build visual language models (VLMs) for multi-modal tasks. Many efforts (Dai et al., 2024; Liu et al., 2024b; Lin et al., 2023) in this field have achieved remarkable performance on visual language understanding. In these works, visual inputs are projected onto LLMs' semantic space through a vision model like CLIP (Radford et al., 2021) to bridge two modalities by including text-image alignment objectives.

In addition to visual understanding, another essential research direction in combining visual and language modalities is visual generation. There are two popular approaches for text-guided image generation. One approach employs diffusion models (Rombach et al., 2022a), a powerful tool for various generation tasks. The other line of work converts visual content into discrete tokens through vector quantization (VQ) and then leveraging autoregressive transformers for high-quality and diverse generation (Esser et al., 2021; Yu et al., 2021; Lee et al., 2022; Tian et al., 2024b; Sun et al., 2024).

Witnessing the rapid advancements in both visual understanding and generation, an emerging trend is to unify these techniques into a single multi-modal framework. Prior to VILA-U, there are two main approaches to achieving such unification: (1) One approach (Liu et al., 2024a; Yu et al., 2023a; Xie et al., 2024) utilizes a VQGAN-based (Esser et al., 2021) tokenizer to convert visual inputs into discrete tokens and leverages an autoregressive model for both understanding and generation. However, (Xie et al., 2024) has shown that visual tokens from VQGAN-based encoder lack semantic information and usually results in a severe performance drop in downstream visual understanding tasks. (2) Another approach (Zhan et al., 2024; Ge et al., 2023b; Jin et al., 2023) utilizes a codebook to quantize features produced by a pre-trained vision model like CLIP. Since CLIP features encode rich semantic information, these approaches generally achieve significantly better performance on understanding tasks. However, these tokenizers lack decoding capability, requiring an external visual generation model, such as a diffusion model, to use the generated visual tokens as conditions for producing visual outputs. This approach adds complexity to infrastructure design. Available large-scale foundation model training pipelines and deployment systems have already been highly optimized for language modeling with next-token prediction. Designing and maintaining an additional stack to support diffusion models would incur significant engineering costs.

In this work, we present **VILA-U**, an *end-to-end autoregressive* framework with a unified next-token prediction objective for both visual and text inputs that can achieve competitive performance on both visual language understanding and generation tasks, without the help of external components like diffusion models. We identify two critical principles to unify vision and language modalities: (1) Existing unified end-to-end autoregressive VLMs cannot achieve competitive visual understanding performance because the discrete VQGAN tokens are trained solely on image reconstruction loss and are not aligned with textual inputs. Therefore, it is crucial to introduce text alignment during VQ vision tower pretraining to enhance perception capabilities. (2) Autoregressive image generation can attain similar quality as diffusion models if trained on high-quality data with sufficient size. Guided by these insights, VILA-U features a unified foundation vision tower that converts visual inputs into discrete tokens through vector quantization and aligns these tokens with textual inputs using contrastive learning. The multi-modal training of VILA-U takes advantage of a unified next-token prediction objective for both visual and textual tokens on a small-size high-quality image-text corpus.

We evaluate VILA-U on common visual language tasks, including image-language understanding, video-language understanding, image generation and video generation. VILA-U significantly narrows the gap in visual understanding performance between end-to-end autoregressive models and continuous-token VLMs, while introducing competitive *native* visual generation capabilities.

2 Related Work

Large Language Models (LLMs). LLMs based on pre-trained large-scale transformers (Vaswani et al., 2017) has drastically revolutionized natural language processing field. Featuring gigantic model size and pre-training data corpus, LLM has achieved remarkable performance on various linguistic tasks. The development of open-source LLMs such as LLaMA (Touvron et al., 2023a), Mixtral (Jiang et al., 2024) and Vicuna (Chiang et al., 2023) has furthered nourished research on how to adopt LLM for complex language tasks. Besides excellent zero-shot generalizability to diverse domains, LLM is commonly finetuned on custom datasets for better performance on specific tasks. Instruction tuning (OpenAI, 2023; Chung et al., 2024; Ouyang et al., 2022) also stands as a key step for better outputs in applying LLMs. In this work, we adopt the LLaMA-2-7B(Touvron et al., 2023a) model as our basic LLM.

Visual Language Models (VLMs). Combining computer vision and natural language processing gives rise to VLM in this LLM era. In VLMs, researchers leverage vision foundation models such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022a) and CoCa (Yu et al., 2022) to extract visual features, align with texts, and feed them into LLM to achieve the cross-modality understanding between texts and visual content. Building upon such progress, many VLMs (Alayrac et al., 2022; Li et al., 2023b; Liu et al., 2024b; Lin et al., 2023; Luo et al., 2024; Tian et al., 2024a) have been designed and trained on extensive vision-language data to achieve remarkable performance on visual understanding and reasoning tasks. VLMs are divided into two types. (1) *BLIP-style* VLMs (Awadalla et al., 2023; Alayrac et al., 2022; Li et al., 2022b; 2023c; Dai et al., 2023; Hong et al., 2023) utilizes cross attention mechanism to fuse language and visual information and optionally apply perceivers (Jaegle et al., 2021) to downsample visual tokens. (2) *LLaVA-style* VLMs (Liu et al., 2023b; Driess et al., 2023; Chen et al., 2023b; AI, 2023; Zhu et al., 2023; Ye et al., 2023; Bai et al., 2023; Aiello et al., 2023; Chen et al., 2023c; Liu et al., 2023a; Lin et al., 2023; Zhang et al., 2023) converts visual inputs to tokens (patches) and pass them through ViTs. The output of ViTs undergoes MLP layers and gets aligned to the language space. In this work, we aim to develop a VLM with visual understanding capacities comparable to prior works, while also possessing the new capacity of visual generation.

Unified Visual Language Models. Numerous efforts have been made to develop unified visual language models capable of generating both text and visual content, including images and videos. There are two mainstream methods to generate visual content in VLMs. Many works (Sun et al., 2023b;a; Jin et al., 2023; Ge et al., 2023b; Li et al., 2023d; Ge et al., 2024; Jin et al., 2024; Ge et al., 2023a) combine VLMs with diffusion models like Stable Diffusion (Rombach et al., 2022a) for high-quality image generation. Other works (Liu et al., 2024a; Yu et al., 2023a; Lu et al., 2023; Team, 2024; Xie et al., 2024) adopt VQGAN-based vision encoders to convert visual inputs into discrete tokens and make LLMs learn to predict them. In this work, we design our framework based on the autoregressive next-token prediction method for visual generation and make our VLM learn to generate visual content effectively.

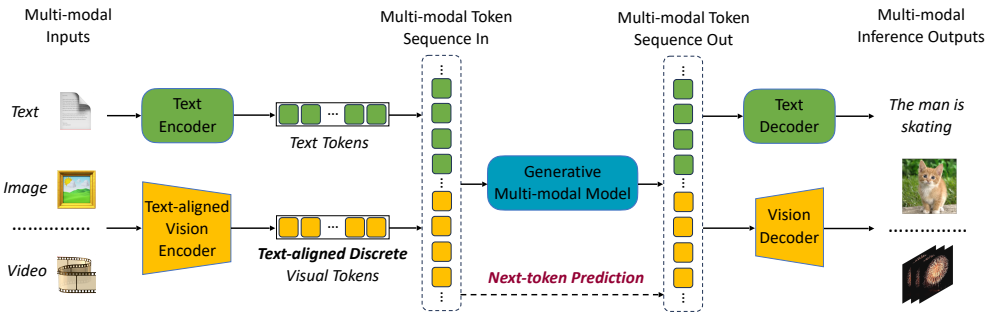

Figure 1: **An overview of our framework's multi-modal training and inference process.** Visual inputs are tokenized into discrete tokens and concatenated with textual tokens to form a multi-modal token sequence. All tokens are involved in our next-token prediction process, enabling a unified training objective. During inference, the output tokens are decoded by our text detokenizer or vision tower decoder to yield multi-modal content.

3 METHODS

This work proposes a multi-modal framework that aims to unify visual and language modalities effectively. The key components enabling such unification are a unified foundation vision tower that converts visual inputs into discrete tokens aligned with text, and a unified multi-modal generative training procedure. An overview of the main multi-modal training and inference process within our framework is depicted in Figure 1.

3.1 UNIFIED FOUNDATION VISION TOWER

To support diverse visual understanding and generation tasks, we first build a unified foundation vision tower to provide appropriate visual features. We propose to include text-image contrastive loss and VQ-based image reconstruction loss in our vision tower training, empowering the text alignment and discrete tokenization abilities for our vision tower. As depicted in Figure 2, the features extracted from images are primarily discretized through residual quantization. Then in one route, the discrete visual features are fed into a decoder to reconstruct the image and compute the reconstruction loss; on the other route, we compute the image-text contrastive loss between the discrete visual features and the textual features provided by a text encoder. With this training procedure, the vision tower learns to extract discrete features suitable for both understanding and generation in our VLM.

Unified Training Recipe. Training the unified vision tower with two objectives from scratch would be difficult. This is because alignment and reconstruction tasks require high-level semantic and low-level appearance features, respectively. Training the entire vision tower from scratch with both objectives could induce conflicting goals. In practice, we observe that training the vector-quantized vision tower from scratch with both image reconstruction and contrastive loss results in a mere 5% Top-1 accuracy for zero-shot image classification on ImageNet (Deng et al., 2009a) after several epochs of training.

To address this issue, we experiment with different training recipes and find the following solution to be most effective. Instead of learning both objectives simultaneously, our training recipe suggests first equipping the model with text-image alignment ability and then learning reconstruction while maintaining alignment ability. We initialize the vision encoder and text encoder with pretrained weights from the CLIP model to ensure good text-image alignment. Next, we freeze the text encoder and keep all vision components trainable using both contrastive and reconstruction loss. The contrastive loss maintains alignment ability, while the reconstruction loss develops reconstruction ability. This training approach converges quickly and yields strong performance. The pre-trained CLIP weights contain learned high-level priors, which are difficult and computationally expensive to learn from scratch. Initializing with these weights enables the binding of low-level and high-level features much faster and more tractably for the vision encoder. With this training recipe, we can train a vision tower that exhibits both good text alignment and image reconstruction abilities. We use weighted sum to combine the text-image contrastive loss and VQ-based image reconstruction loss:

$$\mathcal{L}_{total} = w_{contra}\mathcal{L}_{contra} + w_{recon}\mathcal{L}_{recon} \quad (1)$$

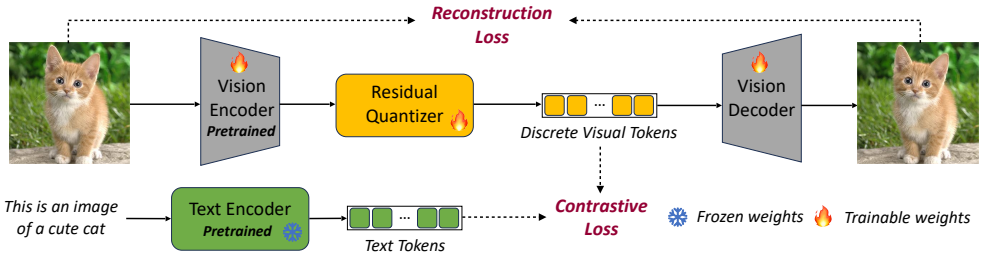

Figure 2: **Overview of our unified foundation vision tower.** Given input images the features extracted by the vision encoder are discretized using residual quantization. Then the discrete vision features are meanwhile put into the vision decoder to reconstruct images and used to perform the text-image alignment. During this process, the reconstruction loss and contrastive loss are computed to update the vision tower, endowing it to produce discrete visual features with text alignment.

In our experiments, we pick w_{contra} = 1 and w_{recon} = 1.

Residual Vector Quantization. Our visual features are discretely quantized, so their representation ability heavily depends on the code size used in our quantizer. Since we hope they contain both high-level and low-level features, we need more capacities in their vector feature space, making a larger code size necessary for good performance in downstream tasks. However, too many codes for each image will result in too many tokens for LLM to produce in the visual generation process, incurring much latency. So in an attempt to increase the vector feature capacity and meanwhile maintain a reasonable number of tokens for LLM, we adopt a residual vector quantization method following RQ-VAE (Lee et al., 2022) to discretize a vector $\mathbf{z}$ as D discrete codes:

$$\mathcal{RQ}(\mathbf{z}; \mathcal{C}, D) = (k_1, \cdots, k_D) \in [K]^D, \tag{2}$$

where $\mathcal{C}$ is the codebook, $K = |\mathcal{C}|$ and k_d is the code of $\mathbf{z}$ at depth d. Starting with $\mathbf{r}_0 = \mathbf{z}$, we recursively perform vector quantization by

$$\begin{aligned} k_d &= \mathcal{Q}\left(\mathbf{r}_{d-1}, \mathcal{C}\right), \\ \mathbf{r}_d &= \mathbf{r}_{d-1} - \mathbf{e}\left(k_d\right), \end{aligned} \tag{3}$$

for each depth $d = 1, 2, \cdots, D$, where $\mathbf{e}$ is the codebook embedding table and $\mathcal{Q}$ is the standard vector quantization:

$$\mathcal{Q}(\mathbf{z}; \mathcal{C}) = \arg\min_{k \in [K]} \|\mathbf{z} - \mathbf{e}(k)\|_2^2. \tag{4}$$

The quantized vector for $\mathbf{z}$ is the sum over the depth dim: $\widehat{\mathbf{z}} = \sum_{i=1}^{D} \mathbf{e}\left(k_i\right)$. Intuitively, in each depth we choose a code to reduce the quantization error. So compared to the standard vector quantization methods, we have D codes to quantize one vector, allowing for finer approximation and larger feature space. During multi-modal training and inference, LLM only needs to predict the code embedding, with codes in different depth sequentially produced by a depth transformer taking the code embedding as the initial input, as we will introduce in Section 3.2. So with this residual quantization, we can enhance the representation capability of our vision tower while incurring little latency.

3.2 Unified Multi-modal Generative Pre-training

Figure 1 presents an overview of our unified multi-modal pre-training process. Our vision tower encoder processes visual inputs sequentially, generating a 1D token sequence. This sequence is then concatenated with text tokens to form a multi-modal sequence. To distinguish between modalities and enable visual content generation, we insert special tokens: `<image_start>` and `<image_end>` at the start and end of image tokens, and `<video_start>` and `<video_end>` at the start and end of video tokens. Video tokens are the direct concatenation of multi-frame image tokens.

Pre-training data form. In terms of unified pre-training data, we leverage different concatenation forms between text and visual tokens to facilitate both understanding and generation. We use `[image, text]`, `[text, image]`, and `[text, video]` forms, with supervision loss added only on the latter modality in each pair to avoid unconditional content generation and promote modality alignment. We also employ an interleaved text and image concatenation form for enhanced

understanding, with supervision loss applied solely to the text. Notably, we exclude the `[video, text]` form during pre-training for efficiency reasons, as we find incorporating it during supervised fine-tuning effectively yields excellent video understanding ability.

Training Objective. Since both visual tokens and text tokens are discrete, we can train our LLM with the general language modeling next-token prediction objective. However, due to the use of residual quantization for visual tokens, the training objectives for text and visual tokens differ slightly. For text tokens, the negative log-likelihood loss is calculated as

$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{T} \log P_\theta \left(y_i | y_{<i} \right), \tag{5}$$

where T is the length of the multi-modal sequence and i only counts when the text token appears at position i. For visual tokens, residual quantization introduces a depth-stacked structure of codes at each visual position j. To address this, we leverage the depth transformer introduced in RQ-VAE (Lee et al., 2022). Specifically, given the code embedding h_j generated by the LLM for visual tokens at position j, the depth transformer autoregressively predicts D residual tokens (k_{j1}, ..., k_{jD}). During training, the input of the depth transformer v_{jd} at depth d is defined as the sum of the code embeddings of up to depth $d-1$ for $d > 1$ such that

$$v_{jd} = \sum_{d'=1}^{d-1} \mathbf{e}(k_{jd'}), \tag{6}$$

and $v_{j1} = h_j$. Thus, the depth transformer predicts the next code for a finer estimation of the feature $\hat{\boldsymbol{z}}_j$ based on the previous estimations up to $d-1$. Then the negative log-likelihood loss for visual tokens is

$$\mathcal{L}_{\text{visual}} = -\sum_{j=1}^{T} \sum_{d=1}^{D} \log P_\delta \left(k_{jd} | k_{j,<d} \right), \tag{7}$$

where T is the length of the multi-modal sequence and j only counts when a visual token appears at position j. During the multi-modal pre-training, the weights of the depth transformer are randomly initialized and updated together with the LLM.

4 Experiments

In this section, we introduce comprehensive experiments to evaluate our method on various visual understanding and generation tasks. Firstly, we outline our experimental setup, including the model architecture, training datasets, and evaluation benchmarks. Subsequently, we evaluate the performance of our unified foundation vision tower. Then, we compare our method with other popular VLMs on various visual understanding and generation benchmarks. Finally, we give some qualitative results.

4.1 Experimental Setup

In our experiments, we employ LLaMA-2-7B (Touvron et al., 2023b) as our base language model. For the vision tower, we choose SigLIP-Large-patch16-256 / SigLIP-SO400M-patch14-384 (Zhai et al., 2023) as our vision encoder architecture, and adopt the residual quantizer, depth transformer as well as the decoder architecture from RQ-VAE (Lee et al., 2022). The quantizer codebook size is 16384. All images and videos are resized to a resolution of 256×256 / 384×384, with each image or video frame converted into a $16 \times 16 \times 4$ / $27 \times 27 \times 16$ code with the residual depth $D = 4$ / $D = 16$. We train our vision tower on COYO-700M (Byeon et al., 2022) and evaluate it for zero-shot classification and reconstruction performance on ImageNet (Deng et al., 2009b). For visual understanding, we leverage 1M `[image, text]` data from ShareGPT4V (Chen et al., 2023a), 6M interleaved text and image data from MMC4 (Zhu et al., 2024). For visual generation, we incorporate 15M high-quality `[text, image]` data curated from our internal dataset and 1M `[text, video]` data from OpenVid (Nan et al., 2024) datasets. Classifier-free guidance (Ho & Salimans, 2022) is employed for visual generation with a CFG value of 3.

For examining visual understanding ability, we evaluate our model on the widely adopted zero-shot image-based visual-language benchmarks including VQAv2 (Goyal et al., 2017), GQA (Hudson &

Manning, 2019), TextVQA (Singh et al., 2019), POPE (Li et al., 2023e), MME (Fu et al., 2024), SEED (Li et al., 2023a), MM-Vet (Yu et al., 2023b) and video-based visual-language benchmarks including ActivityNet (Caba Heilbron et al., 2015), MSVD (Chen & Dolan, 2011), MSRVTT (Xu et al., 2017), TGIF (Li et al., 2016).

To evaluate the visual generation capability, we use MJHQ-30K (Li et al., 2024) and GenAI-Bench (Lin et al., 2024) as our benchmarks. The former adopts the FID between generated images and 30K high-quality images to reflect the overall capability of image generation. The latter is a challenging image-to-text generation benchmark that reflects the comprehensive generative abilities of visual generation models. This benchmark is divided into two categories of prompts: *basic* skills, which include attribute, scene, and relation understanding in text inputs, and *advanced* skills, which encompass counting, differentiation, comparison, and logical relation understanding in text inputs.

4.2 Unified Foundation Vision Tower

We present the commonly used metrics reconstruction FID (rFID) and Top-1 accuracy for zero-shot image classification on ImageNet to measure the reconstruction and text alignment capabilities of the unified foundation vision tower in Table 1. Please refer to the Appendix A.1 for the qualitative reconstruction results. Our model achieves significantly better reconstruction results than VQ-GAN. Our rFID is slightly inferior to that of RQ-VAE when using the same code shape. This is expected as the introduction of contrastive loss during training, aimed at enhancing image understanding, led to a decrease in reconstruction quality. For the text alignment capability, our unified vision tower achieves a Top-1 accuracy of 73.3 / 78.0 under 256 / 384 resolution. This demonstrates the exceptional text alignment capability of our unified vision tower. However, it is worth noting that both the rFID and Top-1 accuracy of the vision tower only serves as a medium indicator. As the unified vision tower is an integral component of the entire autoregressive model, we believe that its performance on downstream tasks, such as visual understanding and generation, holds greater significance.

Table 1: The reconstruction FID (rFID) and Top-1 accuracy for zero-shot image classification of our unified vision tower on ImageNet.

Model	Pretrained Weights	Resolution	Shape of Code	rFID↓	Top-1 Accuracy↑
VQ-GAN	–	256×256	16×16	4.98	–
RQ-VAE	–	256×256	$8 \times 8 \times 4$	3.20	–
RQ-VAE	–	256×256	$16 \times 16 \times 4$	1.30	–
Ours	SigLIP-Large	256×256	$16 \times 16 \times 4$	1.80	73.3
Ours	SigLIP-SO400M	384×384	$27 \times 27 \times 16$	1.25	78.0

4.3 Quantitative Evaluation

Visual Understanding Tasks. Table 2 and Table 3 summarize the comparison between our method and other leading VLMs on the image-language and video-language benchmarks respectively. Compared to the mainstream choice of continuous visual tokens produced by foundation models like CLIP, the VQGAN-based discrete visual tokens have less alignment with text, thus harming VLMs' performance on visual understanding tasks. With our unified foundation vision tower, our model can have a performance close to leading VLMs even with discrete visual tokens.

Visual Generation Tasks. As shown in Table 4, VILA-U can achieve a better FID than other autoregressive methods and have comparable performance with some diffusion based methods. This result shows the feasibility of our method for visual generation. Table 5 summarizes the quantitative results of our method and other visual generation methods on GenAI-Bench. Although Our method is inferior to diffusion-based visual generation methods that have been trained on billions-level image-text pairs,

Method	Type	#Images	FID↓
SD v2.1	Diffusion	–	26.96
SD-XL	Diffusion	2000M	9.55
PixArt	Diffusion	25M	6.14
Playground v2.5	Diffusion	–	4.48
LWM	Autoregressive	–	17.77
Show-o	Autoregressive	36M	15.18
Ours (256)	Autoregressive	15M	12.81
Ours (384)	Autoregressive	15M	7.69

Table 4: Comparison with other visual generation methods on MJHQ-30K evaluation benchmark.

Table 2: Comparison with leading methods on image-based visual language benchmarks. Our performance is close to leading VLMs, surpassing many methods by a large margin under the same LLM size, even with a discrete visual token type. * indicates that images in the training split of these datasets are observed during VLM training.

Method	LLM	Visual Token	Res.	VQAv2	GQA	TextVQA	POPE	MME	SEED	MM-Vet
LLaVA-1.5	Vicuna-1.5-7B	Continuous	336	78.5*	62.0*	58.2	85.9	1510.7	58.6	30.5
VILA	LLaMA-2-7B	Continuous	336	79.9*	62.3*	64.4	85.5	1533.0	61.1	34.9
Unified-IO 2	6.8B from scratch	Continuous	384	79.4*	–	–	87.7	–	61.8	–
InstructBLIP	Vicuna-7B	Continuous	224	–	49.2	50.1	–	–	53.4	26.2
IDEFICS-9B	LLaMA-7B	Continuous	224	50.9	38.4	25.9	–	–	–	–
Emu	LLaMA-13B	Continuous	224	52.0	–	–	–	–	–	–
LaVIT	LLaMA-7B	Continuous	224	66.0	46.8	–	–	–	–	–
DreamLLM	Vicuna-7B	Continuous	224	72.9*	–	41.8	–	–	–	36.6
Video-LaVIT	LLaMA-2-7B	Continuous	224	80.2*	63.6*	–	–	1581.5	64.4	35.0
Emu2-Chat	Emu2-37B	Continuous	448	84.9*	65.1*	66.6*	–	–	–	–
MM-Interleaved	Vicuna-13B	Continuous	224	80.2*	60.5*	61.0	–	–	–	–
DEEM	Vicuna-7B	Continuous	448	68.2*	55.7*	–	–	–	–	37.4
CM3Leon-7B	7B from scratch	Discrete	256	47.6	–	–	–	–	–	–
LWM	LLaMA-2-7B	Discrete	256	55.8	44.8	18.8	75.2	–	–	9.6
Show-o	Phi-1.5-1.3B	Discrete	256	59.3*	48.7*	–	73.8	948.4	–	–
SEED-LLaMA	Vicuna-7B	Discrete	224	66.2	–	–	–	–	51.5	–
Ours	LLaMA-2-7B	Discrete	256	75.3*	58.3*	48.3	83.9	1336.2	56.3	27.7
Ours	LLaMA-2-7B	Discrete	384	79.4*	60.8*	60.8	85.8	1401.8	59.0	33.5

Table 3: Comparison with leading methods on video-based visual language benchmarks. The performance of our method is close to state-of-the-art VLMs, surpassing many methods under the same LLM size, even with a discrete visual token type.

Method	LLM	Visual Token	Res.	MSVD-QA	MSRVTT-QA	TGIF-QA	Activity Net-QA
Unified-IO 2	6.8B from scratch	Continuous	384	52.1	42.5	–	–
Emu	LLaMA-13B	Continuous	224	–	18.8	8.3	–
VideoChat	Vicuna-7B	Continuous	224	56.3	45	34.4	–
Video-LLaMA	LLaMA-2-7B	Continuous	224	51.6	29.6	–	–
Video-ChatGPT	LLaMA-2-7B	Continuous	224	64.9	49.3	51.4	35.2
Video-LLava	Vicuna-7B	Continuous	224	70.7	59.2	70.0	45.3
Video-LaVIT	LLaMA-2-7B	Continuous	224	73.5	59.5	–	50.2
Emu2-Chat	Emu2-37B	Continuous	448	49.0	31.4	–	–
LWM	LLaMA-2-7B	Discrete	256	55.9	44.1	40.9	–
SEED-LLaMA	Vicuna-7B	Discrete	224	40.9	30.8	–	–
Ours	LLaMA-2-7B	Discrete	256	73.4	58.9	51.3	51.6
Ours	LLaMA-2-7B	Discrete	384	75.3	60.0	51.9	52.7

our method has comparable performance with SD v2.1 (Rombach et al., 2022b) and SD-XL (Podell et al., 2023) on *advanced* prompts even trained with magnitude-level less data. This further shows that VILA-U can learn the correlation among visual and textual modalities effectively with our unified training framework. For video generation, we evaluate our method on VBench Huang et al. (2024) and compare it against Open-Sora Zheng et al., CogVideo Hong et al. (2022), and CogVideoX Yang et al. (2024). The results, presented in Table 6, demonstrate that our method achieves performance that is better than CogVideo and comparable to Open-Sora, highlighting the effectiveness of our approach.

4.4 Qualitative Evaluation

Visual Understanding. To validate the effectiveness of VILA-U in comprehensive visual understanding tasks, we apply it in several understanding and reasoning tasks, as some examples shown in Figure 3 and Figure 4. From the results, we can see the versatility of VILA-U in various tasks including visual captioning and visual question answering. Besides, our model has inherited some important capabilities from VILA (Lin et al.,

Method	Total Score↑	Quality Score↑	Semantic Score↑
Open-Sora	75.91	78.82	64.28
CogVideo	67.01	72.06	46.83
CogVideoX	81.61	82.75	77.04
Ours (256)	74.01	76.26	65.04

Table 6: Comparison with other visual generation methods on VBench (Huang et al., 2024).

Table 5: Comparison with other visual generation methods on GenAI-Bench (Lin et al., 2024). The results show that our method outperforms previous autoregressive visual generation methods. For *advanced* prompts that require better text following ability to generate, our method can have a relatively small performance gap with diffusion-based methods, even with much less training data.

Method	Type	#Training Images	Attribute↑	Scene↑	Relation↑			Overall↑
					Spatial	Action	Part	
SD v2.1	Diffusion	2000M	0.80	0.79	0.76	0.77	0.80	0.78
SD-XL	Diffusion	2000M	0.84	0.84	0.82	0.83	0.89	0.83
Midjourney v6	Diffusion	–	0.88	0.87	0.87	0.87	0.91	0.87
DALL-E 3	Diffusion	–	0.91	0.90	0.92	0.89	0.91	0.90
LWM	Autoregressive	–	0.63	0.62	0.65	0.63	0.70	0.63
Show-o	Autoregressive	36M	0.72	0.72	0.70	0.70	0.75	0.70
Ours (256)	Autoregressive	15M	0.78	0.78	0.77	0.78	0.79	0.76
Ours (384)	Autoregressive	15M	0.75	0.76	0.75	0.73	0.75	0.73

(a) VQAScores on *basic* prompts of GenAI-Bench

Method	Type	#Training Images	Count↑	Differ↑	Compare↑	Logical↑		Overall↑
						Negate	Universal	
SD v2.1	Diffusion	2000M	0.68	0.70	0.68	0.54	0.64	0.62
SD-XL	Diffusion	2000M	0.71	0.73	0.69	0.50	0.66	0.63
Midjourney v6	Diffusion	–	0.78	0.78	0.79	0.50	0.76	0.69
DALL-E 3	Diffusion	–	0.82	0.78	0.82	0.48	0.80	0.70
LWM	Autoregressive	–	0.59	0.58	0.54	0.49	0.52	0.53
Show-o	Autoregressive	36M	0.70	0.62	0.71	0.51	0.65	0.60
Ours (256)	Autoregressive	15M	0.70	0.71	0.74	0.53	0.66	0.64
Ours (384)	Autoregressive	15M	0.68	0.67	0.71	0.51	0.64	0.61

(b) VQAScores on *advanced* prompts of GenAI-Bench

2023) including multi-image understanding, in-context learning, as shown in Figure 5 and Figure 6. More visualizations can be found in the Appendix A.2.

Visual Generation. We present some examples of the visual generation results in Figure 7. Our model can be employed in both image generation and video generation, even trained with a relatively small data corpus. In the given examples, our method can generate nice-looking images and continuous videos adhering to the user's input. More visualizations can be found in the appendix A.3.

5 Ablation Study

5.1 Impact of Contrastive Loss to Visual Understanding

We include contrastive loss in vision tower training, which endows it with the text alignment ability. During our multi-modal training, such text alignment ability is crucial in enhancing modality fusion and performance on downstream visual language tasks. We validate the importance of this alignment by training the vision tower with and without the contrastive loss, evaluating its impact on visual language understanding performance. For this ablation study, we randomly sample 25M data from

Prompt: `<video>\n`
Elaborate on the visual and narrative elements of the video in detail.
VILA-U(Ours): The video features a person wearing a blue t-shirt, khaki pants, and a cap, executing a series of breakdancing moves on a concrete road. The background includes a clear sky, a bridge, and a train passing by. The dancer's movements are dynamic, with spins, flips, and handstands. The lighting is bright, suggesting daytime, and the shadows indicate the sun is high. The dancer's attire is casual, and the setting appears to be a quiet road with minimal traffic.

Figure 3: VILA-U can correctly caption videos and cover all the details, thanks to the text alignment of our vision encoder.

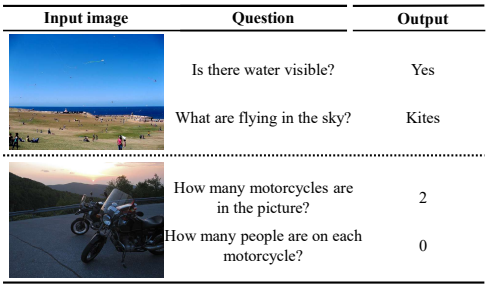

Figure 4: VILA-U has good visual question answering capability. The images and questions are from the test split of VQAv2 dataset.

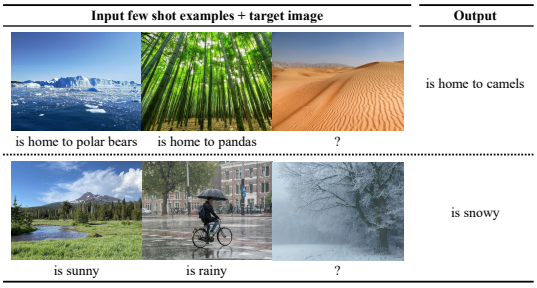

Figure 5: VILA-U has good in-context learning capability. We feed two image-text pairs and a third image as the context to prompt the VLM.

Prompt: Image 1: `<img1>\n` Image 2: `<img2>\n`
What is the common thing and the difference between image 1 and image 2?
VILA-U(Ours): The common thing in both images is the presence of a panda bear. However, the difference between image 1 and image 2 is that the panda bear in image 1 is lying down on a rock, while the panda bear in image 2 is lying down on a log.

Prompt: Image 1: `<img1>\n` Image 2: `<img2>\n`
What is the common thing and the difference between image 1 and image 2?
VILA-U(Ours): The common thing in both images is the presence of trees, but the difference between the two images is the angle of the sunlight. In image 1, the sun is shining through the trees, creating a bright and lively atmosphere. In contrast, the sun is not visible in image 2, which shows a darker and more subdued forest scene.

Figure 6: VILA-U can correctly reason over multiple images.

COYO-700M to train the vision tower. For multi-modal training, we use ShareGPT4V and MMC4 without text-image and text-video data. The results of the first two lines in Table 7 demonstrate the crucial role of text alignment in achieving strong visual language understanding performance. Scaling the dataset size from 25M to 700M further enhances performance, highlighting the importance of learning text alignment on a large-scale dataset.

Table 7: Impact of contrastive loss to visual understanding.

Pretrained Weights	Data size	Loss Type	Top-1 Accuracy	VQAv2	POPE	MME	SEED	MM-Vet
SigLIP-Large	25M	Recon.	–	57.7	75.1	937.7	38.7	15.3
SigLIP-Large	25M	Recon. + Contra.	62.9	68.0	83.7	1219	50.4	20.8
SigLIP-Large	700M	Recon. + Contra.	73.3	75.3	83.9	1336.2	56.3	27.7

5.2 Impact of Contrastive Loss to Visual Generation

We conduct two experiments to demonstrate the influence of contrastive loss to generation performance. For efficiency, we conduct only text-to-image pretraining and utilize Sheared-LLaMA-1.3B (Xia et al., 2023) instead of LLaMA-2-7B as the LLM. In the first experiment, we use the RQ-VAE as the vision tower, which has an rFID of 1.30. In the second experiment, we employ our unified vision tower. Results are shown in Table 8. Our Unified Vision Tower yielded slightly worse FID results than the RQ-VAE on MJHQ-30K, possibly due to its inferior rFID resulting from the contrastive loss.

Table 8: Impact of contrastive loss to visual generation.

Vision Tower	LLM	Resolution	rFID ↓	FID ↓
RQ-VAE (Lee et al., 2022)	Sheared-LLaMA-1.3B	256 × 256	1.30	12.0
Ours	Sheared-LLaMA-1.3B	256 × 256	1.80	13.2

Table 9: Impact of CFG.

CFG Value	FID ↓
1.0	14.1
2.0	13.0
3.0	12.8
5.0	13.2

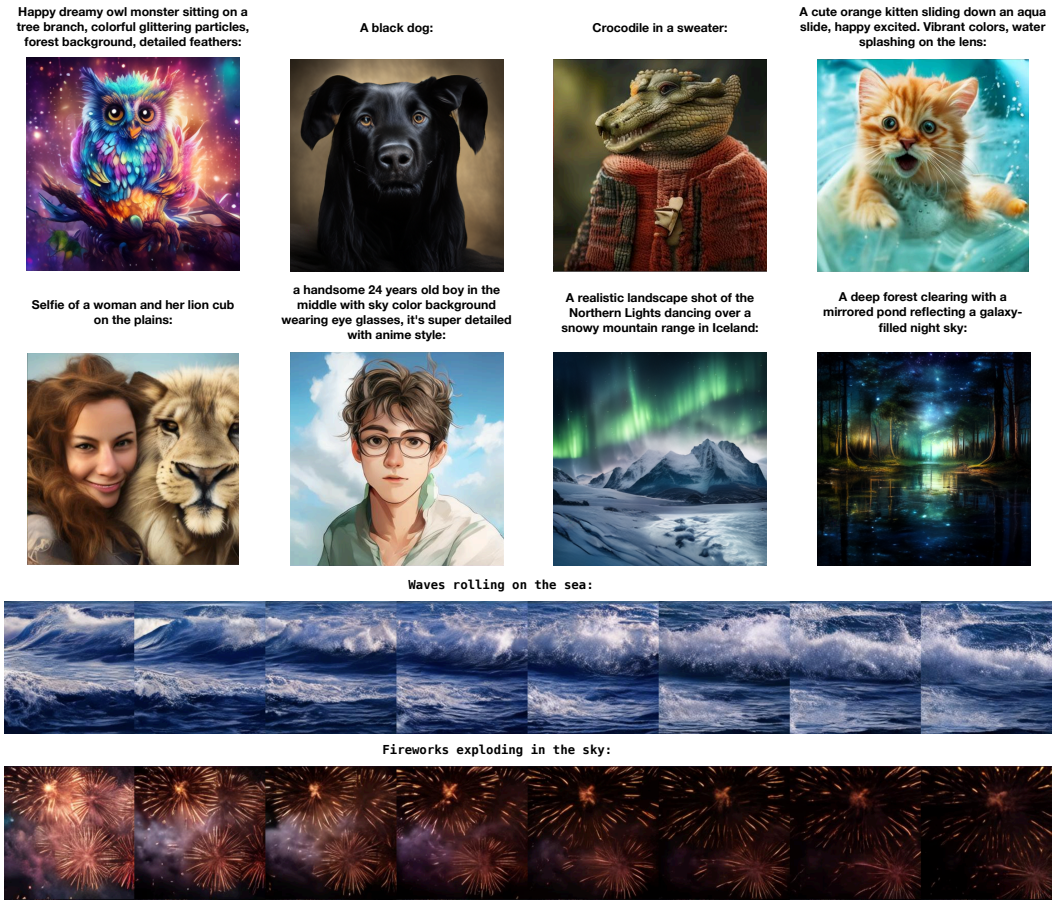

Figure 7: VILA-U can generate high-quality images and videos given text input.

5.3 Impact of Classifier-free Guidance

We adopt classifier-free guidance during the visual content generation. We investigate the impact of the CFG value on our 256-resolution model. Results presented in Table 9 indicate that a CFG value of 3.0 yields the best FID score.

6 Conclusion and Limitation

We present VILA-U, a novel and unified visual language model that integrates video, image and language understanding and generation tasks into one autoregressive next-token prediction framework. Our method is not only more concise than most VLMs that leverage additional components like diffusion models for unifying visual generation and understanding, but also demonstrates that autoregressive methods can achieve comparable performance to state-of-the-art VLMs. We believe VILA-U can serve as a general-purpose framework for diverse visual language tasks.

As demonstrated in Section 5.2, the introduction of contrastive loss impacts the reconstruction ability of the vision tower. Balancing these two capabilities within the unified vision tower presents an interesting and complex challenge that requires further exploration. Additionally, we currently do not observe significant synergy or mutual enhancement between understanding and generation tasks. In the future, we aim to investigate and explore more effective methods to enable these tasks to complement and reinforce each other, thereby fully realizing the untapped potential of a unified visual language model.

REFERENCES

ADEPT AI. Fuyu-8B: A multimodal architecture for AI agents. https://www.adept.ai/blog/fuyu-8b, 2023.

Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. Jointly training large autoregressive multimodal models. *arXiv preprint arXiv:2309.15564*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL https://doi.org/10.5281/zenodo.7733589.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. Technical report, Alibaba Group, 2023. https://arxiv.org/abs/2303.08774.

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.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023a.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023c.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. URL https://api.semanticscholar.org/CorpusID:258615266.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, pp. 248–255. Ieee, 2009a.

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*, pp. 248–255. Ieee, 2009b.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023a.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. In *The Twelfth International Conference on Learning Representations*, 2023b.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

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*, pp. 6700–6709, 2019.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv:2401.04088*, 2024.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, CHEN Bin, Chengru Song, Di ZHANG, Wenwu Ou, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In *The Twelfth International Conference on Learning Representations*, 2023.

Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

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

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022b.

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*, pp. 19730–19742. PMLR, 2023b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*, 2023d.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023e.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4641–4650, 2016.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, et al. Deem: Diffusion models serve as the eyes of large language models for image perception. *arXiv preprint arXiv:2405.15232*, 2024.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

OpenAI. Chatgpt. `https://openai.com/blog/chatgpt/`, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

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

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

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024. URL `https://arxiv.org/abs/2406.06525`.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023a.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023b.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024a.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024b. URL `https://arxiv.org/abs/2404.02905`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023a.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023.

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.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023a.

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*, 2023b.

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

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, march 2024. *URL https://github. com/hpcaitech/Open-Sora*, 1(3):4.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024.

APPENDIX

A QUALITATIVE RESULTS

A.1 RECONSTRUCTION

Figure 8: Visualization of the reconstruction results from text-aligned discrete visual tokens.

We present qualitative reconstruction results in Figure 8 for our 256 / 384 resolution vision tower. These vision towers effectively reconstruct images in detail using text-aligned discrete visual tokens.

A.2 VISUAL UNDERSTANDING

Input image	Question	Output
	What animals are shown here?	Zebras
	What color are they?	Black and white
	What is the color of the far right board?	Blue
	How many men are in the picture?	2
	Is she cooking bread?	Yes
	What room is this?	Kitchen
	Did the man hit the bat?	No
	Is the man's legs together or apart?	Apart

Figure 9: Image understanding results. Examples are taken from the test split of VQAv2 dataset.

Prompt: <video>\n What do the drops do 5 times ?
VILA-U(Ours): The drops fall down the window 5 times, and each time they hit the window, they create a new set of drops.

Prompt: <video>\n how many guys are laughing ?
VILA-U(Ours): 2 guys are laughing.

Prompt: <video>\n where is the man running and then he fall ?
VILA-U(Ours): The man is running on a treadmill in a gym, and then he falls off the treadmill.

Prompt: <video>\n how many boy holds up the candy bar and opens mouth ?
VILA-U(Ours): The video shows one boy holding up the candy bar and opening his mouth.

Figure 10: Video understanding results. Examples are taken from the test split of TGIF dataset.

We provide more image understanding and video understanding examples in Figure 9 and Figure 10. VILA-U successfully answers the questions accurately.

A.3 Visual Generation

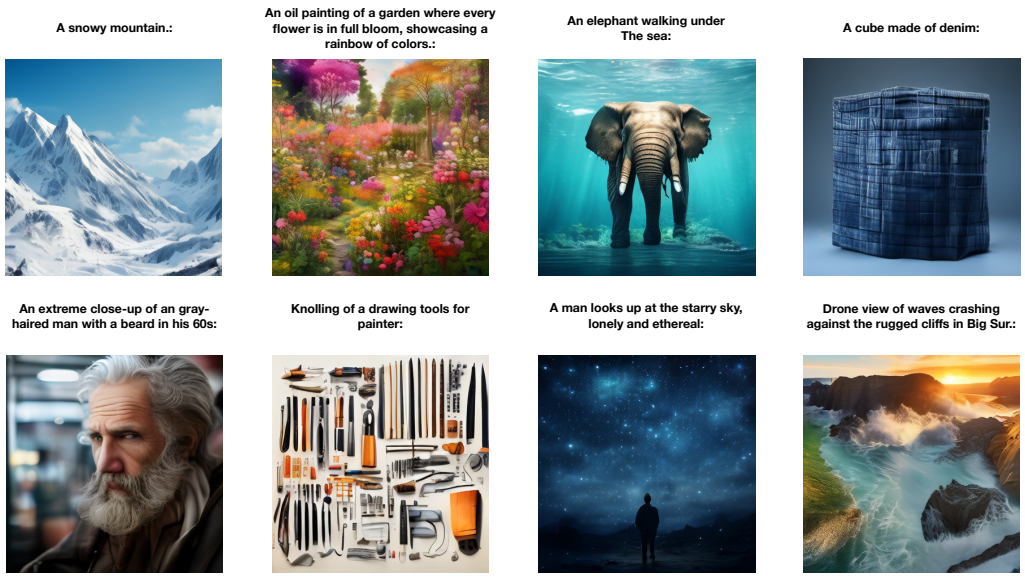

Figure 11: Image generation results. VILA-U can generate high-quality images given text input.

Figure 12: Video generation results. VILA-U can generate high-quality videos given text input.

We provide more image generation and video generation examples in Figure 11 and Figure 12. VILA-U can generate high-quality images and videos given text input.

A.4 In-context Learning Examples

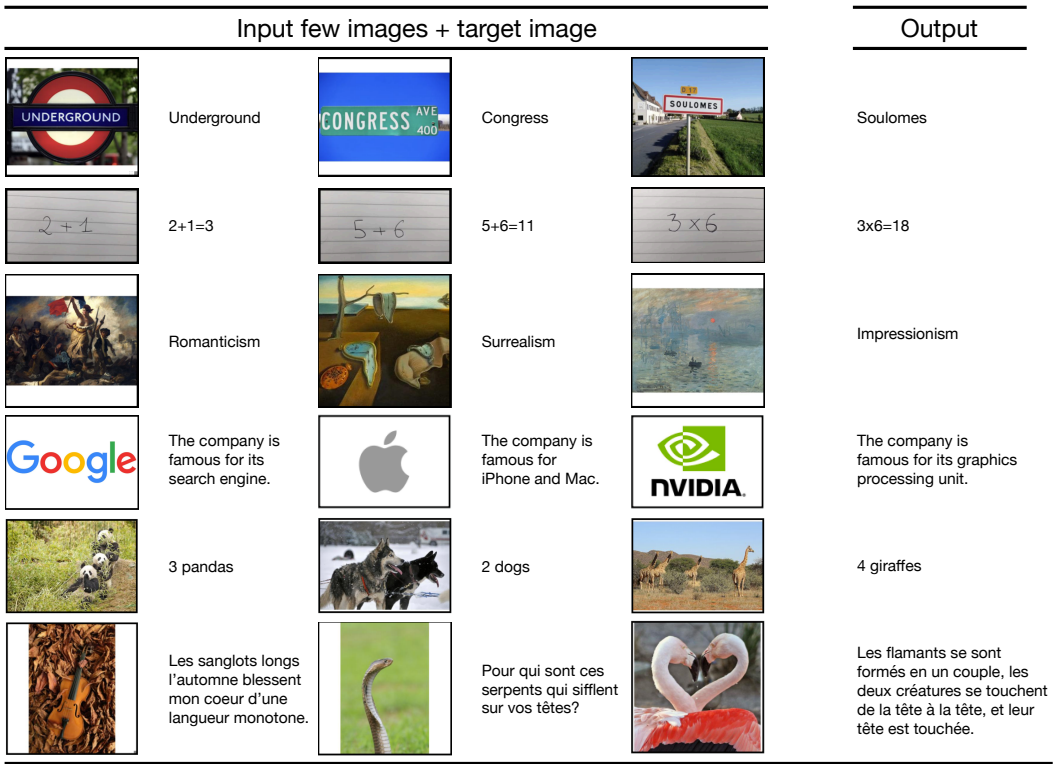

Figure 13: In-context learning examples. We try all in-context learning examples in Lin et al. (2023). The results demonstrate that VILA-U has inherited good in-context learning capabilties.

We provide more qualitative results to demonstrate in-context learning capabilities of VILA-U in Figure 13. VILA-U exhibits good in-context learning capabilties.

B Difference with related works

Prior to VILA-U, unified visual language models were dominated by two mainstream approaches:

(1) Represented by LWM, CM3Leon and Show-o which utilizes a VQGAN-based tokenizer to convert visual inputs into discrete tokens. However, as these tokenizers are trained solely with a reconstruction objective, the resulting tokens lack rich semantic information. This limitation leads to poor performance on multimodal understanding tasks. But it can easily support autoregressive visual generation and the generated visual tokens can be seamlessly decoded into visual outputs using the lightweight decoder of VQGAN.

(2) Represented by AnyGPT SEED-LLaMa and LaViT, which utilizes a codebook to quantize features produced by a pre-trained ViT model like CLIP. Since CLIP features encode rich semantic information, these approaches generally achieve significantly better performance on understanding tasks compared to VQGAN-based tokenizers. However, these tokenizers lack decoding capability, requiring an external visual generation model, such as a diffusion model, to use the generated visual tokens as conditions for producing visual outputs.

Compared to these two mainstream approaches, VILA-U introduces a solution that addresses the limitations of both. We design a unified vision tower that extracts features with rich semantic information, similar to CLIP, while also supporting image reconstruction capabilities akin to VQGAN. This is achieved by incorporating both reconstruction loss and contrastive loss into the autoencoder training process, along with utilizing residual quantization to enhance the representation capability of the visual features. Building on this foundation, we develop a single end-to-end autoregressive framework that eliminates the need for external visual generation models required by approach 2 and significantly outperforms the understanding results of methods in approach 1.

C Failed Training Recipes.

We experiment with numerous training recipes and find none to be as effective as our final approach. We list four alternative recipes and discuss their shortcomings compared to our final recipe: 1) Load pre-trained CLIP weights into the text encoder only; 2) Load pre-trained RQ-VAE weights for the vision encoder and decoder while training other parts from scratch; 3) Freeze the vision encoder; 4) Make the text encoder trainable.

Recipes 1) and 2) fail due to the lack of pre-trained CLIP weights for the vision encoder. Training a CLIP model from scratch typically requires numerous GPU days with a large global batch size (e.g., 32k). However, VQ-based reconstruction training necessitates a relatively small global batch size (e.g., 512) for steady improvement. With such a small batch size, training a text-aligned vision tower from scratch would be prohibitively time-consuming and resource-intensive.

Recipe 3) fails because freezing the vision encoder prevents it from learning the low-level features essential for reconstruction. In this case, the burden of reconstruction falls entirely on the vision decoder, but it is impossible to reconstruct images well using only semantic features.

Recipe 4) fails because the quantized features are chaotic during the initial training steps, and the contrastive loss disrupts the text encoder weights, slowing down the entire training process.

In contrast, our final training recipe leverages pre-trained CLIP weights for the vision encoder, enabling it to maintain learned semantic features rather than grasping them from scratch. This allows us to train with a small batch size while keeping the vision encoder trainable, facilitating the learning of low-level features for reconstruction during training.