

UniCompress: Token Compression for Unified Vision–Language Understanding and Generation

Ziyao Wang^{1,2} Chen Chen¹ Jingtao Li¹ Weiming Zhuang¹ Jiabo Huang¹
 Ang Li² Lingjuan Lyu^{1,*}

¹Sony AI ²University of Maryland, College Park

Abstract

Unified models aim to support both understanding and generation by encoding images into discrete tokens and processing them alongside text within a single autoregressive framework. This unified design offers architectural simplicity and cross-modal synergy, which facilitates shared parameterization, consistent training objectives, and seamless transfer between modalities. However, the large number of visual tokens required by such models introduces substantial computation and memory overhead, and this inefficiency directly hinders deployment in resource constrained scenarios such as embodied AI systems. In this work, we propose a unified token compression algorithm UNICOMPRESS that significantly reduces visual token count while preserving performance on both image understanding and generation tasks. Our method introduces a plug-in compression and decompression mechanism guided with learnable global meta tokens. The framework is lightweight and modular, enabling efficient integration into existing models without full retraining. Experimental results show that our approach reduces image tokens by up to 4 $\times$, achieves substantial gains in inference latency and training cost, and incurs only minimal performance degradation, which demonstrates the promise of token-efficient unified modeling for real world multimodal applications.

1. Introduction

Recent research on multimodal learning has been moving towards unified models that can *understand* and *generate* images within a single autoregressive framework [24, 34, 42]. A common solution is to encode images into discrete visual tokens produced by a learned tokenizer, then feeds these tokens, together with text tokens, into a language-model backbone. This shared token space enables a broad spectrum of multimodal tasks (e.g., image captioning [4], VQA [39], image editing [3]) to be handled by one archi-

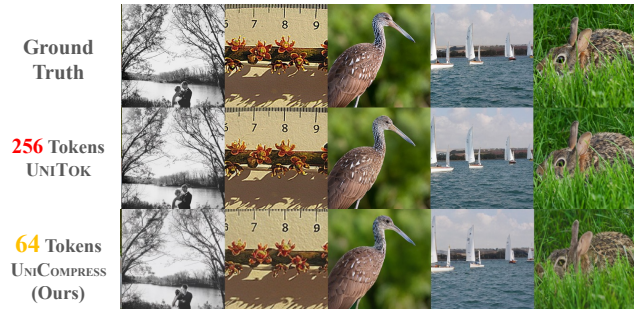


Figure 1. We propose UNICOMPRESS, a plug-in-and-play token compression algorithm for unified models. The samples are from UniTok [24].

ture, simplifying deployment and multi-task training.

However, a practical limitation of these unified models is *token efficiency*. Tokenizers from the discrete codebook family (e.g., VQ-VAE [27], the dVAE used in DALL·E [26], and VQGAN [9]) often map a 512×512 image to $32 \times 32 = 1024$ tokens (i.e. downsampling by a factor of 16 along height and width). Unified models must then use long visual sequences for understanding and adopt equally long sequences for generation, increasing memory footprint, training cost, and inference latency. Sharing a single visual tokenizer across understanding and generation [16, 24] reduces engineering complexity but does not reduce the sequence-length. A straightforward solution is to compress the visual tokens. However, experimental results demonstrate that **naïve downsampling or uniform token pruning, while effective for image understanding, significantly degrades generation performance by more than 15%**. This sets the main challenge since image generation relies on fine-grained, spatially consistent tokens to accurately reconstruct details, hence is sensitive to pruning.

Another challenge is ensuring low-cost. While we can train another tokenizer with better token efficiency (e.g. one-D tokenizer) to replace the existing one, but doing so often requires downstream language model training from scratch, which is costly. A more practical approach should

*Correspondence to: Lingjuan Lyu <Lingjuan.lv@sony.com>

be modular, allowing seamless integration with existing tokenizers without the need for full retraining. Thus, the secondary challenge is *how to develop a plugin-based compression method compatible with any tokenizer, avoiding expensive LLM retraining.*

To address the challenges, we introduce UNICOMPRESS, a plug-in token compression framework that reduces visual tokens by up to $4\times$ while maintaining quality for both understanding and generation. Motivated by [12, 15], UNICOMPRESS inserts two lightweight modules around an off-the-shelf discrete tokenizer: (i) a *compressor* that converts a dense $H \times W$ token grid into a short sequence of compressed tokens *augmented by a small set of learnable global meta tokens* capturing holistic semantics; and (ii) a *global-guided decompressor* that reconstructs high-fidelity token grids conditioned on those global tokens, restoring long-range structure. We first train the tokenizer together with the compressor and decompressor for image reconstruction task, then freeze the compressed tokenizer and lightly fine-tune the language model on the compressed tokens for both understanding and generation. The result is a more efficient unified model with more compact input/output sequence, while retaining the understanding and generation performance through global-guided decompression.¹ Intuitively, the compressed tokens carry salient local evidence, whereas the global meta tokens provide scene-level constraints. During generation, the decompressor uses these global tokens as semantic anchors to autoregressively refine local textures and boundaries, mitigating the detail loss observed with uniform token compression.

On standard understanding and generation benchmarks, UNICOMPRESS reduces visual tokens by $4\times$ (for example, $256 \rightarrow 64$) while keeping performance within small margins (≤ 3 -pt drop on understanding; ≤ 5 -pt FID increase on generation). Relative to uncompressed baselines, UNICOMPRESS yields up to 41.8% lower inference latency and 15.4% shorter training time, together with substantial latency savings from shorter sequences. These results indicate a practical path to unified models suitable for resource-constrained platforms. Our key contributions are as follows:

- We highlight token efficiency as a bottleneck in unified models and show that naïve token compression disproportionately harms *generation*. We formalize the objective of a single compact visual token space usable for both understanding and generation.
- We propose UNICOMPRESS, a plug-in compression framework with *global-guided* autoregressive decompression. It shortens the visual sequence while preserving generation detail and seamlessly integrates into existing unified models.
- We demonstrate strong empirical results across different

¹We instantiate a fixed compression ratio for controlled comparisons; the design naturally extends to content-adaptive rates (see §3).

unified models. Our method achieves up to a $4\times$ token reduction while maintaining competitive performance on both understanding and generation tasks. For both understanding and generation tasks, we cap the performance drop at $\leq 5\%$, and on some benchmarks we fully match pre-compression performance.

2. Related Work

Unified Foundation Model. Recent advances in multi-model learning have led to vision foundation models [41] and unified multimodal models [24, 34] that support both visual understanding and image generation within a single framework. These models typically encode images into discrete token sequences and process them alongside text using a language model backbone. DREAMLLM [8] treats image and text as a joint sequence and leverages autoregressive modeling to seamlessly compose multimodal content. VILA-U [34] moves away from diffusion-based generation, opting instead for a fully token-level decoder that unifies captioning and image synthesis via next-token prediction. To address the limitations of low-capacity tokenizers, UNITOK [24] expands the expressiveness of visual tokens using a multi-codebook quantizer, balancing semantic abstraction and reconstruction detail. VARGPT [42] takes a hierarchical approach by predicting both content and resolution scale, enabling controllable and efficient image generation over multiple granularities. Meanwhile, UNIFORK [21] questions the viability of fully shared backbones and proposes a Y-shaped design that splits deeper layers by task to reduce interference while preserving early fusion. Other works utilize diffusion model as the generation head. For instance, **OpenUni** [33] and **Bagel** [7] further push token-level autoregressive generation with diffusion models. To facilitate joint training and architectural simplicity, many of these models adopt a shared tokenizer across modalities, often using ViT-style patch embeddings or VQ-based discrete representations (e.g., dVAE or VQ-GAN) [35, 39].

However, representing each image with hundreds or even thousands of tokens introduces substantial computational overhead. This limitation highlights the need for a unified compression framework that can reduce token redundancy while preserving task performance.

Visual Token Compression of Foundation Model. A key bottleneck in implementing VLMs lies in the large number of discrete tokens used to represent images. Each image is typically encoded into hundreds or even thousands of tokens, which significantly increases sequence length and leads to high memory usage, latency, and computational costs—particularly during inference. Current token compression works mainly focus on image understanding tasks. [1, 20, 30, 32, 36] prune image tokens either before feeding them into the LLM or within the LLM layers based

on attention scores. These understanding-oriented pruning methods aim to reduce pipeline FLOPs without fine-tuning, thereby improving inference efficiency. For image generation models [28], MaskGIT [5] reduces inference latency through masked token prediction. [31] improves image generation efficiency by reformulating autoregressive decoding as next-scale prediction instead of next-token prediction. HMAR [17] further combines next-scale prediction with masked generation to enhance decoding speed. However, inference-time pruning or prediction optimization techniques are typically task-specific and cannot be directly applied to both understanding and generation tasks in a unified setting [40]. Other works such as TiTok [37] uses a 1D tokenizer to compress visual tokens. While this works well for understanding tasks, it underperforms on generation tasks due to the loss of spatial information. This highlights the need for an efficient unified token compression framework.

3. Our Method: UNICOMPRESS

3.1. Overview

As shown in Fig. 2, UNICOMPRESS augments the visual tokenizer with three lightweight modules while keeping the LLM unchanged: a global-token extractor that uses one-way cross-attention to summarize scene-level semantics; a pooling-based compressor that reshapes the token grid and aggregates non-overlapping patches (e.g., 2×2 , 4×4) into a shorter sequence and an autoregressive decompressor that later expands the compact representation back to a dense token grid for the image decoder. During inference, for image understanding tasks, we place the visual subsequence into the LLM input (after a linear projection to the LLM embedding space). For image generation, the LLM predicts global tokens and compressed visual tokens, we then feed them to the codebook and decompressor to recover the final image. Training is done in two stages: we first train the tokenizer with UNICOMPRESS modules with a reconstruction loss; then we freeze the compressed tokenizer and mildly finetune the LLM model. This “compress once, reuse for both” interface reduces sequence length and compute while preserving global structure. Moreover, our approach readily adapts to other unified model designs, including configurations with multiple image tokenizers or diffusion models.

3.2. Global Token Extraction via Cross-Attention

Let the base encoder output a continuous sequence of visual tokens $\mathbf{X} \in \mathbb{R}^{T \times d}$ where $T = H \times W$ and d is the embedding size. We introduce a small set of learnable meta query tokens $\mathbf{Q} \in \mathbb{R}^{N_g \times d}$ that extract global context from $\mathbf{X}$ using one-way cross-attention: meta tokens query the full image-token field, while image tokens pass through unchanged. For each image, the image-specific global tokens

$\mathbf{G} \in \mathbb{R}^{N_g \times d}$ are computed as

$$\mathbf{G} = \text{MHA}(\mathbf{Q}W_Q, \mathbf{X}W_K, \mathbf{X}W_V), \quad (1)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are learned projections and MHA denotes multi-head attention. We apply residual and normalization on the meta branch,

$$\mathbf{G} \leftarrow \text{LN}(\mathbf{Q} + \mathbf{G}). \quad (2)$$

In practice, N_g is much smaller than T , so the additional sequence length is minor while providing strong global guidance for layout and object relations. Global tokens use their own learned positional embeddings.

3.3. Image Token Compression via Average Pooling

To shorten the visual sequence, we aggregate local tokens within non-overlapping spatial patches. We first reshape $\mathbf{X}$ back to its $H \times W$ grid using the tokenizer’s canonical rasterization. The default operation is fixed-size average pooling applied on the embedding field, which reduces spatial redundancy while preserving coarse structure. Given a downsampling factor s , the compressed continuous sequence is

$$\hat{\mathbf{X}}^{\text{cont}} = \text{AvgPool}(\mathbf{X}, s), \quad \tilde{T} = T/s^2. \quad (3)$$

We then insert the compressed visual segment into the multimodal sequence using three learned special embeddings: an image-begin token [IMG_BOS], a separator token [IMG_SEP] that splits global and local tokens, and an image-end token [IMG_EOS]. For understanding tasks (e.g., image captioning, VQA), depending on the unified model’s design, the language model consumes either continuous tokens or discrete tokens.

For the generation task, the tokenizer includes a discrete codebook of size K . We use the original quantizer to quantize compressed global and local tokens for image generation training. The indices have spatial size $H/s \times W/s$ and take values in $\{1, \dots, K\}$. We use the same codebook for global and local streams:

$$\hat{\mathbf{Z}}^{(g)} \in \{1, \dots, K\}^{N_g}, \quad \hat{\mathbf{Z}}^{(x)} \in \{1, \dots, K_x\}^{\tilde{T}}. \quad (4)$$

This dual representation allows the same compression mechanism to support continuous features for understanding and discrete targets for generation. Image and text tokens share positional embeddings. The target sequence is ordered as

$$[\text{IMG_BOS}], \hat{\mathbf{Z}}_{1:N_g}^{(g)}, [\text{IMG_SEP}], \hat{\mathbf{Z}}_{1:\tilde{T}}^{(x)}, [\text{IMG_EOS}]. \quad (5)$$

3.4. Autoregressive Decompression Guided by Global Tokens

In generation, the language model autoregressively outputs both the global meta tokens and the compressed local tokens

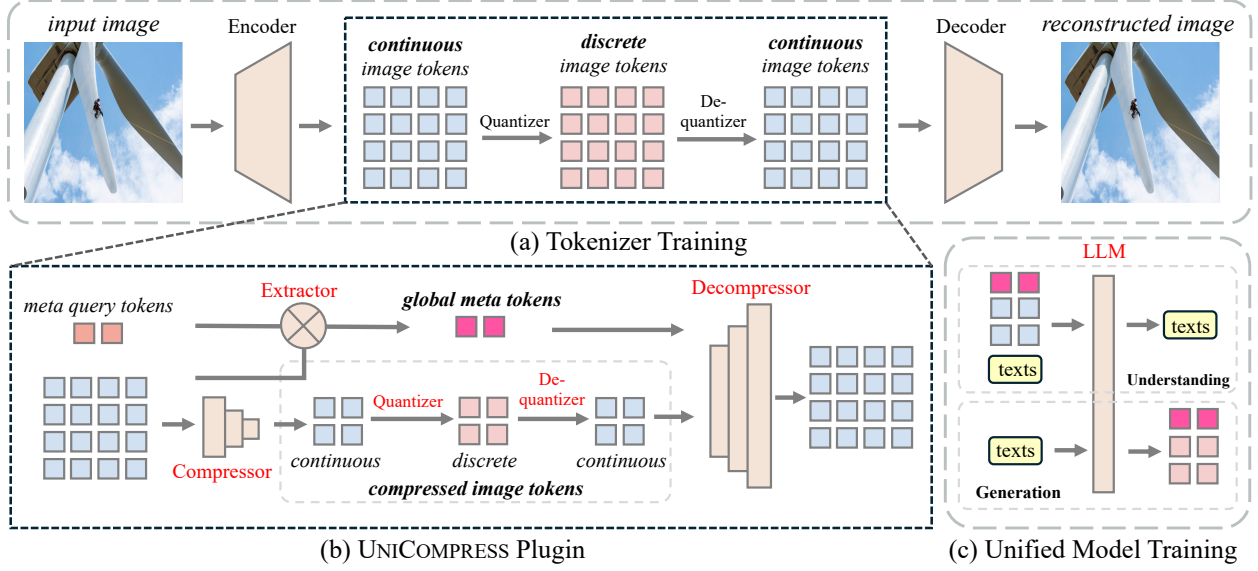


Figure 2. Overview of UNICOMPRESS. The tokenizer is augmented with three modules: a global token extractor, a token compressor, and an autoregressive decompressor. The language model consumes a compact visual sequence for understanding and produces compressed-domain targets for generation.

in the discrete domain. These indices are mapped back to continuous compressed features by codebook lookups,

$$\hat{\mathbf{G}} = \mathcal{E}(\hat{\mathbf{Z}}^{(g)}), \quad \hat{\mathbf{X}}^{\text{deq}} = \mathcal{E}(\hat{\mathbf{Z}}^{(x)}), \quad (6)$$

where $\mathcal{E}$ denotes the codebook.

Given $(\hat{\mathbf{G}}, \hat{\mathbf{X}}^{\text{deq}})$, the decompressor expands the compact representation into a dense sequence of continuous tokens at the original resolution expected by the image decoder. We implement f_{dec} as a Transformer decoder with masked self-attention over the generated dense prefix and cross-attention to $(\hat{\mathbf{G}}, \hat{\mathbf{X}}^{\text{deq}})$ at every layer. At raster step t , the next dense token $\mathbf{x}_t$ is predicted by

$$\mathbf{x}_t = f_{\text{dec}}(\mathbf{X}_{<t}^{\text{dense}}, \hat{\mathbf{X}}^{\text{deq}}, \hat{\mathbf{G}}), \quad (7)$$

where $\mathbf{X}_{<t}^{\text{dense}}$ are previously generated dense tokens and causal masking is applied along the generation order.

Training uses teacher forcing against the tokenizer’s dense targets $\mathbf{X}$. The reconstruction objective combines a token-level regression term in the dense feature space with a codebook consistency term:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{reg}} + \lambda_{\text{cb}} \mathcal{L}_{\text{cb}}. \quad (8)$$

3.5. Lightweight Training Pipeline

The overall procedure is modular and keeps changes on the tokenizer side.

Stage one (tokenizer-side training). We freeze the LLM and train the tokenizer stack with the reconstruction

objective $\mathcal{L}_{\text{recon}}$. The stack includes: the learnable meta-query global extractor (producing $\mathbf{G}$), the fixed average-pooling compressor (producing $\hat{\mathbf{X}}^{\text{cont}}$), the codebook $\mathcal{E}$, and the decompressor f_{dec} . This stage learns to map a dense sequence $\mathbf{X}$ to a compact pair $(\mathbf{G}, \hat{\mathbf{X}}^{\text{cont}})$ and back to dense tokens with high fidelity.

Stage two (LLM training). We freeze the tokenizer and train the LLM on compact data. For understanding, the LLM consumes continuous tokens $\{\mathbf{G}, \hat{\mathbf{X}}^{\text{cont}}\}$. For generation, the LLM autoregressively outputs the discrete indices for both streams; the discrete tokens are then de-quantized via $\mathcal{E}_g, \mathcal{E}_x$ to $(\hat{\mathbf{G}}, \hat{\mathbf{X}}^{\text{deq}})$ and expanded by f_{dec} to dense tokens for the image decoder. Because the LLM interface is a standard autoregressive sequence over special tokens and visual indices, UNICOMPRESS integrates into existing unified model backbones without architectural modification.

4. Experiments

We evaluate our unified token compression framework on a wide range of multimodal understanding and generation tasks. The experiments are designed to answer the following questions:

- Q1** Can our method preserve vision-language understanding accuracy under visual token compression?
- Q2** Does the generation performance remain competitive when decoding from compressed image tokens?
- Q3** How much efficiency gain (training time, inference time, and FLOPs) does compression bring?

Table 1. Unified model performance on visual understanding benchmarks (higher is better). XXX-COMPRESSED denotes the same backbone with our plug-in token compression ($s=2$, $N_g=4$). 'MME Cog.' refers to the score of MME Cognition.

Methods	GQA	MME Cog.	MME	POPE	Seed-bench	TextVQA	MMMU	MM-Bench
UNITOK	55.71	251.79	1162.50	82.66	49.22	24.90	26.22	40.34
UNITOK-COMPRESSED	53.07	235.00	1036.05	79.36	48.07	24.66	26.22	42.14
VILA-U	53.43	247.65	1130.20	81.45	47.88	24.11	25.72	39.55
VILA-U-COMPRESSED	51.90	241.24	1105.35	80.75	47.30	23.85	25.55	39.94
VARGPT	58.12	269.30	1290.65	88.04	50.54	26.18	27.13	44.22
VARGPT-COMPRESSED	55.90	265.83	1272.80	84.99	48.41	25.34	27.40	41.15
UNIFORK	54.20	246.51	1120.25	81.22	48.82	24.36	26.22	39.84
UNIFORK-COMPRESSED	50.10	232.86	1045.05	78.97	47.95	24.89	25.90	40.50
OPENUNI	53.33	243.59	1108.70	80.60	48.39	24.30	25.80	39.28
OPENUNI-COMPRESSED	49.80	230.21	1030.10	78.17	47.51	23.10	25.72	39.88
BAGEL	60.05	277.80	1312.40	89.20	51.10	26.90	34.05	45.10
BAGEL-COMPRESSED	59.10	274.50	1304.10	88.60	50.80	36.50	27.80	44.50

Table 2. Performance of the original and compressed unified models on image generation benchmarks. XXX-COMPRESSED inserts UNICOMPRESS without changing the LM interface. Lower FID and higher CLIP indicate better quality.

Methods	FID(↓)	CLIP(↑)
UNITOK	16.14	30.5
UNITOK-COMPRESSED	16.33	25.0
VILA-U	14.80	29.8
VILA-U-COMPRESSED	16.37	28.9
VARGPT	14.77	24.2
VARGPT-COMPRESSED	15.02	21.6
UNIFORK	17.83	25.5
UNIFORK-COMPRESSED	20.24	26.0
OPENUNI	16.45	26.7
OPENUNI-COMPRESSED	24.29	22.3
BAGEL	12.73	32.0
BAGEL-COMPRESSED	17.22	28.8

4.1. Experimental Setup

Models and Datasets. We adopt Llama-3.2-1B [11] as the language backbone. For pre-training, we use lightweight datasets JDB [29] (generation) and ShareGPT4V_PT [6] (understanding), together for a single epoch with accumulated batch size=128 and learning rate=5e-5. We then apply a lightweight, single-epoch fine-tuning on ShareGPT4V for one epoch with accumulated batch size=256 and learning rate=1e-4. Unless otherwise specified, all results are obtained with this training schedule. We use the original tokenizer in the unified model.

Baselines. We compare 6 representative, high-performing unified model backbones and their compressed variants: UNITOK [24], VILA-U [34], VARGPT [42], UNIFORK [21], and OPENUNI [33], and BAGEL [7]. Among these methods, UNITOK, VILA-U, and VARGPT use one

image tokenizer for both understanding and generation tasks, which is the main focus of our UNICOMPRESS. Beyond them, we also plugin-and-play more diverse baselines using UNICOMPRESS, including different tokenizers for understanding and generation (*i.e.* UNIFORK), and unified model with diffusion model (*i.e.* OPENUNI and BAGEL). For each unified model, we create a *compressed* version by inserting our plug-in stack (global extractor, compressor, decompressor) into their vision tokenizer (and diffusion model, if applicable) while keeping the language model interface unchanged. Our setup here uniformly refers to the versions of each unified model that use 256 image tokens (except for OPENUNI and Bagel, since they use diffusion models), and that use Llama-3.2-1B as the language model. Unless stated otherwise, we set the downsampling factor to $s = 2$ (*i.e.*, $4\times$ fewer local tokens) and use $N_g = 4$ global tokens, which our ablation identifies as the best accuracy–efficiency trade-off. There is no prior unified model compression baseline that jointly supports understanding and generation under a single autoregressive interface; therefore we report each unified model against its own *compressed* counterpart.

Benchmarks. We evaluate on standard multimodal benchmarks—GQA [14], MME [10], POPE [22], Seed-bench [18], TextVQA [2], MMMU [38], and MM-Bench [23] using the official splits and metrics. All methods share the same input budgets and decoding hyperparameters. For understanding, the language model directly consumes the continuous visual tokens $\{\mathbf{G}, \hat{\mathbf{X}}\}$ produced by the enhanced tokenizer, without quantization, delimited in the prompt by `[IMG_BOS] ... [IMG_SEP] ... [IMG_EOS]`. We evaluate the generation performance on MJHQ-30K dataset [19]. We report Fréchet Inception Distance (FID) [13], which compares the means and covari-

Table 3. Wall-clock time with/without plug-in token compression. Understanding: ShareGPT4V_PT (train), GQA (inference). Generation: JDB (train), MJHQ-30K (inference). Lower is better. Although the model is trained on the two datasets jointly in other experiments, the training times in this table were measured by training on each dataset separately.

Methods	Und. Train (h)	Gen. Train (h)	Und. Inference (min)	Gen. Inference (min)
UniTok	0.94	4.60	5.41	32.25
UniTok-COMPRESSED	0.81	3.89	5.25	18.96
VILA-U	0.89	4.25	5.12	30.15
VILA-U-COMPRESSED	0.77	3.65	5.09	17.55
VARGPT	1.13	5.52	5.45	38.74
VARGPT-COMPRESSED	0.92	4.67	5.43	22.71
UniFORK	0.90	4.30	5.82	29.90
UniFORK-COMPRESSED	0.77	3.70	5.80	17.60

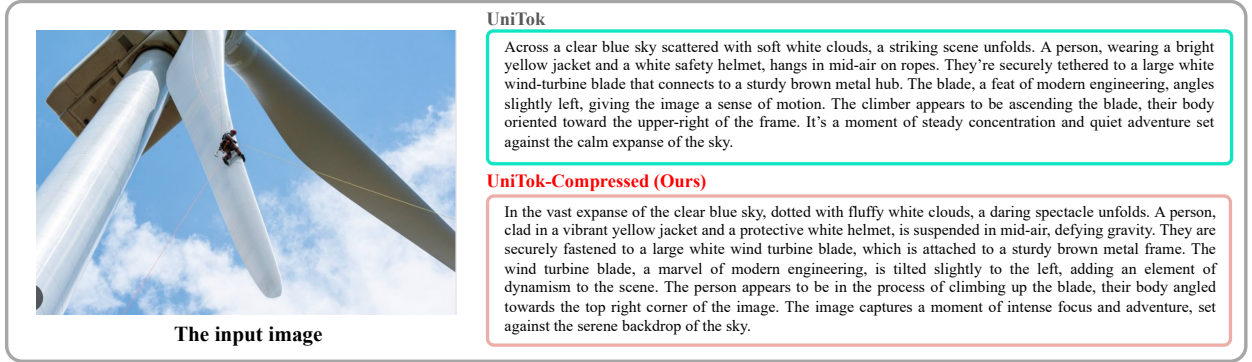


Figure 3. Understanding task examples: generating the texts that describe the image.

ances of Inception-feature distributions between generated and real images; lower values indicate better quality. We also report CLIPScore [25] (from CLIP), the cosine similarity between text and image embeddings in CLIP’s joint space, which measures alignment between images and text; higher values indicate better alignment. For generation, the language model predicts compressed-domain indices for *both* global and local tokens, which are then de-quantized and decompressed into dense features before the image decoder. For methods without a diffusion model (i.e., UniTok, VILA-U, VARGPT, and UniFORK), we use identical sampling procedures and inference budgets (same guidance, number of steps, and temperature) to ensure comparability. For methods with a diffusion model (i.e., OPENUNI and BAGEL), we keep the same downsampling factor $s = 2$, while applying slightly different compression schemes to the diffusion model’s input and output sizes.

4.2. Vision-Language Understanding Performance

Table 1 reports results on standard visual understanding benchmarks, comparing each baseline with its compressed counterpart. Across methods, our compression framework consistently maintains strong accuracy despite substantial reductions in visual token length. For example, UniTok-COMPRESSED shows only minor drops on GQA

(55.71→53.07) and POPE (82.66→79.36), while largely preserving performance on MME and MMMU. For baselines that incorporate diffusion models, UniCOMPRESS likewise preserves understanding quality: Seed-bench within OPENUNI decreases only slightly (48.39→47.51), and OPENUNI-COMPRESSED even outperforms the original on MM-Bench. These findings confirm the robustness of our globally guided decompression design, which enables accurate visual–semantic reconstruction from compact representations. A further advantage of our approach is its modular, plug-in nature: without modifying backbones or introducing additional supervision, the same token compressor can be seamlessly applied across architectures, highlighting its generality for transformer-based vision-language systems.

Figure 3 compares captions produced from the same LLM when using dense UniTok tokens versus our pooled, globally guided compressed tokens. The compressed variant preserves the key entities and relations in the scene (person, wind turbine, blue sky) while maintaining spatial layout and action cues (climbing direction, body orientation). This example illustrates that understanding remains robust under pooling-based token compression, consistent with our quantitative trends.

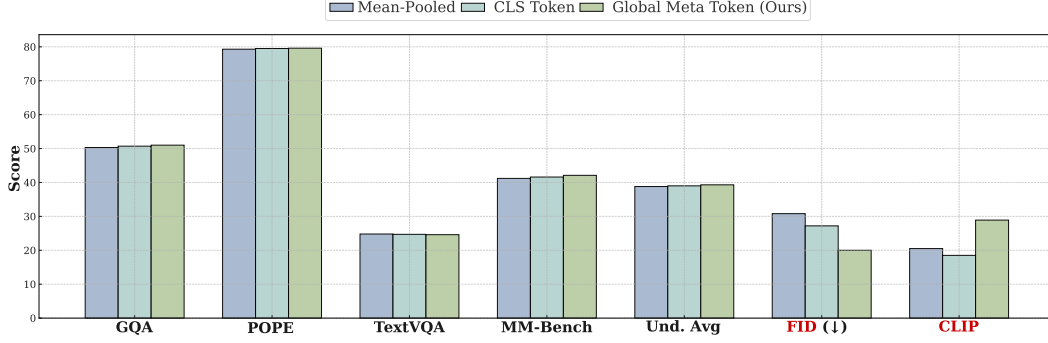


Figure 4. Ablation on global token type. Results use $N_g=4$. Our global meta token yields competitive understanding and notably stronger generation quality (lower FID, higher CLIP).

4.3. Image Generation Performance

Table 2 reports FID (lower is better) and CLIP similarity (higher is better) for each baseline and its compressed variant. Overall, compression tends to increase FID moderately while often reducing CLIP alignment, though the magnitude varies by backbone.

On lighter backbones, the degradation is small: UNiTOK changes from 16.14/30.5 (FID/CLIP) to 16.33/22.0, and VARGPT from 14.77/24.2 to 15.02/21.6. VILA-U is comparatively robust, with FID 14.80→16.37 and CLIP 29.8→28.9. Among stronger models, UNIFORK shows a slight CLIP *increase* (25.5→26.0) despite FID rising to 20.24. BAGEL attains the best baseline quality (12.73 FID, 32.0 CLIP); its compressed version remains competitive at 17.22 FID and 28.8 CLIP. The largest drop is observed for OPENUNI (16.45/26.7→24.29/22.3), indicating greater sensitivity to token reduction for that design. On lighter backbones, the degradation is small: UNiTOK changes from 16.14/30.5 (FID/CLIP) to 16.33/22.0, and VARGPT from 14.77/24.2 to 15.02/21.6. VILA-U is comparatively robust, with FID 14.80→16.37 and CLIP 29.8→28.9. Among stronger models, UNIFORK shows a slight CLIP *increase* (25.5→26.0) despite FID rising to 20.24. BAGEL attains the best baseline quality (12.73 FID, 32.0 CLIP); its compressed version remains competitive at 17.22 FID and 28.8 CLIP. The largest drop is observed for OPENUNI (16.45/26.7→24.29/22.3), indicating greater sensitivity to token reduction for that design.

Taken together, these results show that, while token compression introduces some loss in FID, many models retain strong semantic alignment (CLIP), and several backbones (e.g., VILA-U, VARGPT, UNIFORK) remain close to their full-token counterparts, supporting the viability of compressed visual inputs for high-quality image synthesis.

4.4. Training and Inference Efficiency

Table 3 summarizes the training and inference efficiency across all models. Compressed variants consistently reduce

training time and inference latency, with the largest gain observed in generation settings. In particular, UNiTOK-COMPRESSED reduces generation inference time from 32.25 minutes to 18.96 minutes, corresponding to a relative speedup of over 40%. Similar improvements are observed in VILA-U and VARGPT. These results highlight a major advantage of our approach: it enables end-to-end acceleration of generation pipelines, which has been difficult to achieve in prior token compression frameworks. Most existing methods primarily optimize training throughput or reduce memory usage, but show limited effect on actual decoding time. By contrast, our framework improves runtime efficiency without compromising task performance. The benefits of compression are especially pronounced in autoregressive generation, where each token directly impacts latency. These findings demonstrate the practical utility of our method for real-world deployment, particularly in scenarios with limited compute budgets.

Table 4. Ablation on local token compression (pooling/selection). All rows target the same token budget ($\times 4.0$) as our default setting with $s = 2$. Results use $N_g = 4$.

Compressor	GQA	TextVQA	MM-Bench	Und. Avg	FID(↓)
AvgPool (Ours)	53.07	24.66	42.14	39.29	20.01
MaxPool	50.80	24.40	41.90	39.03	20.15
Strided Conv (2×2)	51.20	24.70	42.10	39.33	19.95
Top- k Token Selection	49.90	24.00	41.20	38.37	21.00
Learned Pooling (Gated)	51.30	24.75	42.20	39.42	19.90

4.5. Ablation Studies

Effect of Token Compression Ratio. We study how the token *keep ratio* ($1/s^2 \in \{1, 1/4, 1/16, 1/64\}$ with stride $s \in \{1, 2, 4, 8\}$) affects performance by compressing the $H \times W$ token grid via non-overlapping average pooling. This design preserves spatial layout, hence only integer pooling windows (e.g., 2×2 , 4×4) are used. As shown in Fig. 6, (1) we use pooling-based compression to retain spatial structure, therefore the ratios are restricted to $1/s^2$ with $s \in \{2, 4, \dots\}$; (2) understanding is relatively robust

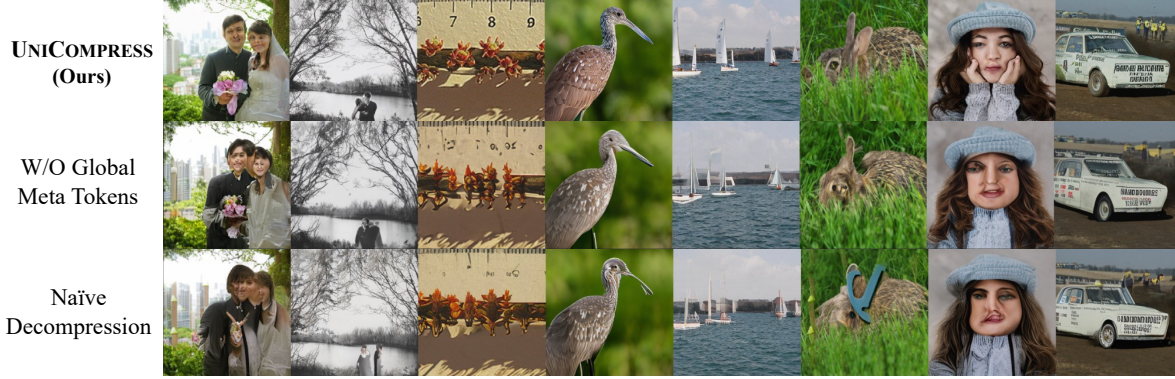


Figure 5. UNICompress preserves the most visual information under compression by using global meta tokens and autoregressive decompressor.

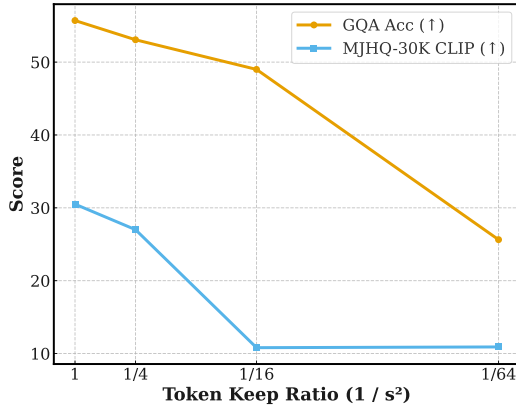


Figure 6. Effect of token keep ratio on accuracy. GQA (understanding) vs. MJHQ-30K CLIP (generation).

to compression (GQA drops moderately from 55.71 at 1 to 49.00 at $1/16$), while generation is far more sensitive (MJHQ-30K CLIP falls sharply from 30.5 to ~ 11), validating our introduction claim; (3) a keep ratio of $1/4$ ($s=2$) provides a good trade-off. Further compression yields noticeable degradation, especially on generation metrics.

Global token type. We compare three ways to form global tokens: mean-pooled image tokens, a CLS token from a ViT encoder, and our learnable global meta tokens. As shown in Fig. 4, understanding metrics (e.g., GQA, TextVQA, MM-Bench, and the averaged score) stay close across all choices, indicating that coarse global summarization is generally sufficient for recognition-style tasks. In contrast, generation quality differs markedly: our global meta tokens achieve substantially lower FID and higher CLIP than mean-pooling or a CLS token. We attribute this gap to the query-based extraction that explicitly “reads” the whole token map and writes image-specific global seman-

tics, which provides stronger conditioning for the decompressor. Hence, while all types are adequate for understanding, the learnable global meta tokens are crucial to preserve fidelity in generation. We also evaluate the impact of global token number in

Figure 5 visualizes the tokenizer compression stage before any LLM decoding. From top to bottom: ours (UNICompress with autoregressive decompressor), ablation without global meta tokens, and naïve decompression (non-autoregressive). Under the same keep ratio, UNICompress retains the most visual detail and global structure; removing global tokens breaks long-range consistency, while dropping the autoregressive decompressor yields over-smoothed textures and artifacts. This confirms that global guidance and autoregressive decompression are both necessary to preserve fidelity under compression.

5. Conclusion

We present UNICompress, a token compression framework for unified models that supports both understanding and generation. The method adds a lightweight, modular compression-decompression mechanism centered on global token guided reconstruction, which reduces the number of visual tokens while preserving task performance. With a two stage training pipeline that keeps the language model unchanged, UNICompress can be integrated into existing systems without full model retraining. Experiments across diverse benchmarks show up to $4\times$ reduction in visual tokens and over 40% faster generation inference, with minimal loss in accuracy or image quality. In contrast to prior approaches that emphasize only training efficiency or understanding tasks, UNICompress delivers consistent gains in both training and inference and for understanding and generation. These results indicate that compact visual representations can enable practical deployment under limited compute and memory, and they point toward scalable unified models that maintain quality while operating at much lower budgets.

References

- [1] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401, 2025. [2](#)
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [5](#)
- [3] Chengyu Bai, Jintao Chen, Xiang Bai, Yilong Chen, Qi She, Ming Lu, and Shanghang Zhang. Uniedit-i: Training-free image editing for unified vlm via iterative understanding, editing and verifying. *arXiv preprint arXiv:2508.03142*, 2025. [1](#)
- [4] Sanghyun Byun, Jung Ick Guack, Mohanad Odema, Baisub Lee, Jacob Song, and Woo Seong Chung. Unifying vision-language latents for zero-label image caption enhancement. *arXiv preprint arXiv:2510.12931*, 2025. [1](#)
- [5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022. [3](#)
- [6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. [5](#)
- [7] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. [2](#), [5](#)
- [8] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. [2](#)
- [9] 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*, pages 12873–12883, 2021. [1](#)
- [10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mmc: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. [5](#)
- [11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. [5](#)
- [12] John Harvill, Ziwei Fan, Hao Wang, Luke Huan, Anoop Deoras, Yizhou Sun, and Hao Ding. Lossless token sequence compression via meta-tokens. *arXiv preprint arXiv:2506.00307*, 2025. [2](#)
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [5](#)
- [14] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. [5](#)
- [15] Wentao Jiang, Jing Zhang, Di Wang, Qiming Zhang, Zeng-mao Wang, and Bo Du. Lemevit: Efficient vision transformer with learnable meta tokens for remote sensing image interpretation. *arXiv preprint arXiv:2405.09789*, 2024. [2](#)
- [16] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3600–3610, 2025. [1](#)
- [17] Hermann Kumbong, Xian Liu, Tsung-Yi Lin, Ming-Yu Liu, Xihui Liu, Ziwei Liu, Daniel Y Fu, Christopher Re, and David W Romero. Hmar: Efficient hierarchical masked autoregressive image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2535–2544, 2025. [3](#)
- [18] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. [5](#)
- [19] 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, 2024. [5](#)
- [20] Kevin Y Li, Sachin Goyal, Joao D Semedo, and J Zico Kolter. Inference optimal vlms need fewer visual tokens and more parameters. *arXiv preprint arXiv:2411.03312*, 2024. [2](#)
- [21] Teng Li, Quanfeng Lu, Lirui Zhao, Hao Li, Xizhou Zhu, Yu Qiao, Jun Zhang, and Wenqi Shao. Unifork: Exploring modality alignment for unified multimodal understanding and generation. *arXiv preprint arXiv:2506.17202*, 2025. [2](#), [5](#)
- [22] 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*, 2023. [5](#)
- [23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. [5](#)
- [24] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025. [1](#), [2](#), [5](#)
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. Pmlr, 2021. 6
- [26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1
- [27] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 1
- [28] Vikash Sehwal, Xianghao Kong, Jingtao Li, Michael Spranger, and Lingjuan Lyu. Stretching each dollar: Diffusion training from scratch on a micro-budget. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28596–28608, 2025. 3
- [29] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678, 2023. 5
- [30] Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. Lvpunring: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. *arXiv preprint arXiv:2501.13652*, 2025. 2
- [31] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 3
- [32] Han Wang, Yuxiang Nie, Yongjie Ye, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. Dynamic-vlm: Simple dynamic visual token compression for videollm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20812–20823, 2025. 2
- [33] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025. 2, 5
- [34] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 1, 2, 5
- [35] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24135–24146, 2025. 2
- [36] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29836–29846, 2025. 2
- [37] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024. 3
- [38] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5
- [39] Jihai Zhang, Tianle Li, Linjie Li, Zhengyuan Yang, and Yu Cheng. Are unified vision-language models necessary: Generalization across understanding and generation. *arXiv preprint arXiv:2505.23043*, 2025. 1, 2
- [40] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 3
- [41] Weiming Zhuang, Chen Chen, Zhizhong Li, Sina Sajadmanesh, Jingtao Li, Jiabo Huang, Vikash Sehwal, Vivek Sharma, Hirotaka Shinozaki, Felan Carlo Garcia, et al. Argus: A compact and versatile foundation model for vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4418–4429, 2025. 2
- [42] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*, 2025. 1, 2, 5

A. Additional Experimental Results

System. We conduct our experiments on a single-node Ubuntu 22.04 LTS server equipped with an Intel(R) Xeon(R) Platinum 8468V CPU and $8 \times$ NVIDIA H100 80GB GPUs. Unless otherwise noted, all training and inference run on this machine using multi-GPU data parallelism.

Number of global tokens. As shown in Figure 7, we ablate $N_g \in \{0, 2, 4, 8, 16\}$ and find a clear sweet spot at $N_g = 4$. Removing globals ($N_g = 0$) consistently hurts both understanding and generation (e.g., higher FID ≈ 21.4). Introducing a small set ($N_g = 2$) improves all metrics but still trails larger settings. Performance largely plateaus for $N_g \in \{4, 8\}$: understanding scores at 8 are only marginally above 4 (often within ~ 0.2 – 0.3 absolute), and FID at 4 (≈ 20.01) matches 8 (≈ 20.00). Since N_g increases sequence length and compute roughly linearly, we adopt $N_g = 4$ as the best accuracy–efficiency trade-off: it recovers global semantics sufficiently to guide decompression, reaches near-peak accuracy, and avoids the diminishing returns observed at $N_g \geq 8$.

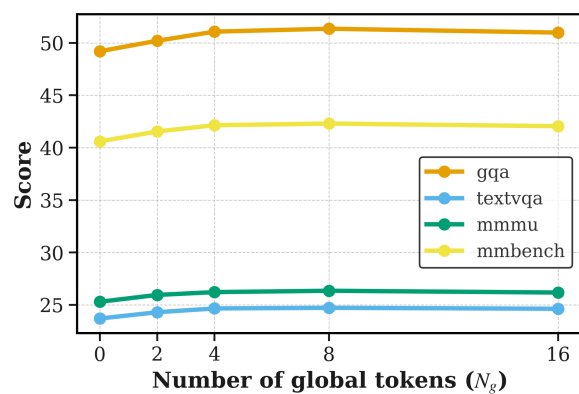
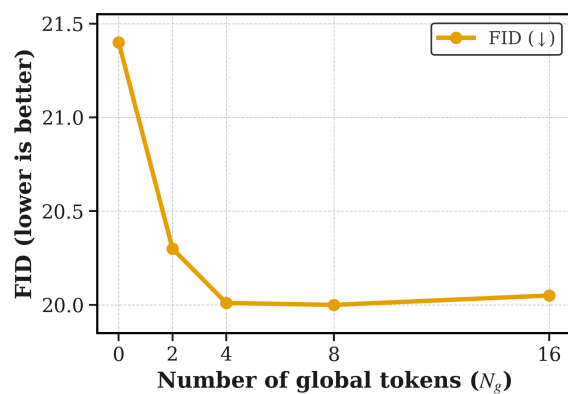


Figure 7. Effect of the number of global tokens N_g . Left: vision-language understanding (GQA, TextVQA, MMMU, MMBench; higher is better). Right: image generation quality (FID; lower is better).