STUDYING IMAGE TOKENIZERS AS VISUAL LANGUAGES IN UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image tokenizers define the "visual language" of unified multimodal models, yet are commonly studied through isolated metrics or generation-/understanding-only evaluations. These evaluations do not fully capture how visual tokens behave when modeled jointly with text. We build a controlled pure-autoregressive testbed and track task-specific validation losses during multimodal continual pretraining across text, image, text-to-image (T2I), and image-to-text (I2T) prediction. We examine how these losses scale and relate to downstream performance, then use them to study multimodal learnability—how well image and text tokens are jointly modeled—and tokenizer design. We find that (1) losses should be analyzed by task, since they exhibit distinct scaling behavior and rank tokenizers differently; and (2) the loss–performance relationship depends on the predicted token space: for a fixed tokenizer, T2I and I2T losses correlate with generation quality, but across tokenizers, the T2I loss–performance relationship shifts with the image-token space, whereas I2T loss, computed over a shared text vocabulary, provides a more consistent signal. I2T loss also correlates with both generation and visual understanding performance after supervised finetuning. Using losses as a lens, we show that (3) better reconstruction does not necessarily yield lower task-specific losses or stronger downstream performance, and that (4) image tokenizer choice can affect text modeling under joint optimization. As case studies, we revisit three tokenizer design axes—the discriminator, semantic supervision, and vocabulary size—to examine their effects on joint modeling and downstream performance. Together, our testbed offers a complementary perspective on image tokenizers as visual languages, highlighting their interplay with text in joint multimodal training.

1 INTRODUCTION

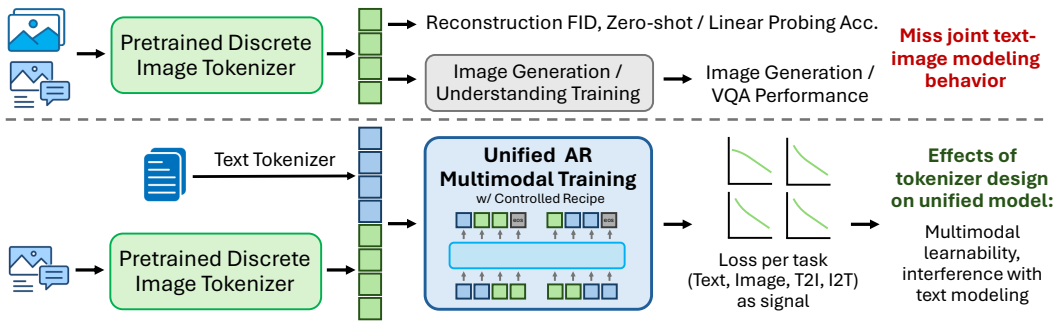

Figure 1: **Studying image tokenizers as visual languages in unified multimodal models.** Prior tokenizer studies focus on downstream-agnostic metrics (reconstruction/probing) or single-axis generation-/understanding-only pipelines, which miss joint text–image modeling behavior. We study tokenizers in the context of unified AR multimodal training, using task-specific losses as a signal to reveal how the tokenizer shapes downstream joint modeling.

Large-scale unified multimodal models have advanced rapidly in both visual generation and understanding by modeling images and text within a single framework. Among these, autoregressive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(AR) modeling is a popular choice due to its compatibility with strong off-the-shelf language models (Team, 2024; Liu et al., 2024; Wu et al., 2025a; Cui et al., 2025). In models that autoregressively predict both text and image tokens, a key component is the discrete image tokenizer, which converts raw pixels into discrete symbols that can be modeled alongside text (Wu et al., 2024; Shi et al., 2025; Ma et al., 2025; Lu et al., 2025). In this sense, an image tokenizer is not merely a preprocessing module; it defines the "visual language" that the language model must learn, align with text, and use for both visual generation and understanding.

However, existing tokenizer studies primarily examine this component outside the unified modeling context, relying on isolated reconstruction-based metrics (e.g., rFID (Heusel et al., 2017)) or ImageNet classification accuracy (Zhao et al., 2025; Zheng et al., 2025a), which do not necessarily translate into better downstream performance (Yao et al., 2025). Others study tokenizers through generation-only or understanding-only downstream pipelines and report benchmark scores (Tong et al., 2024; Yu et al., 2024). These evaluations still leave open how tokenizers affect joint modeling: training with text may change image-token modeling, while image tokenizers may affect cross-modal alignment and text modeling. Tokenizer design choices may therefore be optimized without accounting for their effects under a shared image–text objective.

Motivated by this gap, we study **how the image tokenizer shapes multimodal modeling behavior in unified AR multimodal training**. In Section 3, we develop a controlled, pure-AR continual pretraining recipe using publicly available images and Qwen3 as the language backbone (Yang et al., 2025). Building upon this testbed, we use task-specific validation losses during continual pretraining to characterize tokenizer effects on text, image, text-to-image (T2I), and image-to-text (I2T) modeling. Applying this perspective requires addressing two questions. First, it is unclear whether task-specific losses exhibit similar scaling behavior during mixed image–text AR continual pretraining, or whether they must be interpreted separately. Second, it is unclear how these losses relate to downstream benchmark performance: tokenizers define different visual vocabularies, and next-token prediction loss does not directly measure semantic faithfulness or perceptual quality.

We therefore first establish how to interpret these losses in unified multimodal training (Section 4). All losses decrease with increasing data and model size, but exhibit distinct scaling behavior and rank tokenizers differently across tasks, making an averaged loss too coarse for this analysis. We then relate losses to generation and understanding benchmarks. With a fixed tokenizer, both T2I and I2T losses track generation quality. Across tokenizers, the T2I relationship shifts with vocabulary size and reconstruction fidelity, whereas I2T loss over shared text tokens provides a more consistent signal. I2T loss measured before supervised finetuning (SFT) also correlates with post-SFT generation and general VQA performance, supporting its use as a diagnostic signal.

Using task-specific losses as a lens, we make several observations about how the image tokenizer affects multimodal modeling behavior (Section 5). First, reconstruction fidelity can diverge from *multimodal learnability*—how well image and text tokens are modeled under the shared AR objective—indicating that reconstruction metrics alone are insufficient for evaluating tokenizers in unified multimodal models. Second, tokenizer choice can affect text modeling through the image-token prediction objective under joint training, a cross-modal effect not captured by evaluating generation or understanding alone. We then revisit three tokenizer design axes through ablations and find that (1) a DINO-based discriminator improves reconstruction without consistent downstream gains; (2) semantic supervision strengthens object-level image-token–word associations and learnability despite worse reconstruction; and (3) vocabulary size affects learnability non-monotonically, although a larger vocabulary can improve downstream performance. Together, these results offer a complementary perspective on image tokenizers as visual languages, highlighting their interplay with text in joint multimodal training.

2 PRELIMINARIES

In this section, we review discrete image tokenizers and the key design axes along which they differ. Table 1 summarizes the tokenizers used in our study; further information is provided in Appendix B.

The VQGAN-based image tokenizers studied here (Esser et al., 2021; Yu et al., 2021) follow an encoder-quantizer-decoder architecture. The encoder maps an input RGB image $x \in \mathbb{R}^{H\times W\times 3}$ to a sequence of K continuous vectors $f \in \mathbb{R}^{K\times D}$, where K is the number of image tokens and D is the feature dimension. The quantizer maintains a codebook $Z \in \mathbb{R}^{B\times D}$ of vocabulary size

B and replaces each encoder vector with its nearest codebook entry, yielding quantized vectors $z \in \mathbb{R}^{K \times D}$. The corresponding codebook indices form the discrete image-token sequence modeled by the autoregressive model. The decoder reconstructs an RGB image $\hat{x} \in \mathbb{R}^{H \times W \times 3}$ from z. A standard VQGAN is trained with a vector-quantization (VQ) loss and an autoencoding (AE) loss:

$$\mathcal{L}_{VQ} = ||\text{sg}(f) - z||_2^2 + \beta||f - \text{sg}(z)||_2^2, \quad \mathcal{L}_{AE} = \mathcal{L}_2(x, \hat{x}) + \mathcal{L}_{\mathcal{P}}(x, \hat{x}) + \mathcal{L}_{\mathcal{G}}(x, \hat{x}), \qquad (1)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation, β weights the commitment loss, and $\mathcal{L}_2, \mathcal{L}_{\mathcal{P}}, \mathcal{L}_{\mathcal{G}}$ are the L_2 reconstruction loss, the LPIPS perceptual loss (Zhang et al., 2018), and the adversarial (GAN) loss from the discriminator, respectively.

Architecture. For the encoder/decoder, prior work adopts various architectures, including ConvNet (Esser et al., 2021; Sun et al., 2024), ViT (Yu et al., 2021), and ViT-based models with learnable latent queries (Yu et al., 2024). For codebook design, multi-codebook approaches (Qu et al., 2025; Song et al., 2025; Ma et al., 2025) introduce extra sequence and prediction choices, complicating controlled comparisons; we thus restrict comparisons to single-codebook tokenizers.

Table 1: **Tokenizers studied in our framework:** IBQ (Shi et al., 2025), GigaTok (Xiong et al., 2025), and UniTok (Ma et al., 2025). B: vocabulary size; $\mathcal{L}_{sem}$: semantic loss; Disc.: discriminator type; Usage: codebook usage; rFID: reconstruction FID on ImageNet-1K; †: models trained by us.

Tokenizer	B	$\mathcal{L}_{sem}$	Disc.	Usage	rFID↓
IBQ-1024	1024		PatchGAN	99%	2.24
IBQ-8192	8192		PatchGAN	98%	1.87
IBQ-16384	16384		PatchGAN	96%	1.37
GigaTok	16384	✓	PatchGAN	100%	0.81
GigaTok-DINO	16384	✓	DINO	100%	0.51
UniTok†	16384		DINOv2-S	100%	1.86
UniTok-sem†	16384	✓	DINOv2-S	100%	2.23

Bitwise compression ratio. Bitwise compression depends on resolution $H \times W$, token count K, and vocabulary size B. We fix resolution at 256×256 and $K = 16 \times 16$, varying B within the IBQ family (Shi et al., 2025) in Section 5.4. We defer to future work the study of varying $H \times W$ and K, as changing image resolution alters the visual detail available to the model, while changing K alters sequence length and the image-token budget during training, introducing additional factors into the comparison.

Training objectives. Beyond the standard losses in Eq. (1), we study two design choices in the training objective: **semantic supervision**, using single-codebook UniTok variants without/with CLIP contrastive loss (UniTok/UniTok-sem) (Ma et al., 2025) in Section 5.3; and the **discriminator** used in $\mathcal{L}_G$, using GigaTok variants with two discriminators (GigaTok with PatchGAN and GigaTok-DINO with DINO) (Xiong et al., 2025) in Section 5.1.

3 Framework Construction

In this section, we describe the framework for exploring the image tokenizer's effect on downstream unified multimodal training. In Section 3.1, we introduce the training recipe, which extends a pretrained text language model into a unified autoregressive multimodal model through continual pretraining and supervised finetuning. In Section 3.2, we define the validation loss used to measure modeling quality across tasks. Further training details are provided in Appendix B.

3.1 Training Recipe

Overall setup. We extend the pretrained Qwen3-0.6B, 1.7B, and 4B models (Yang et al., 2025) with B image-token embeddings and corresponding LM-head outputs. New embeddings are initialized from a multivariate normal distribution matching the original embeddings' mean and covariance. We also add special tokens ⟨boi⟩ and ⟨eoi⟩ to delimit flattened image-token sequences. During training, we remove text conditioning from 10% of T2I samples, marked by ⟨unconditional⟩, to enable classifier-free guidance (CFG) (Ho & Salimans, 2022). The resulting model accepts and produces both text tokens and flattened image-token IDs.

Continual pretraining stage. In this stage, we train the unified model on large-scale image-text and pure-text data.

(1) **Data mixture and preprocessing.** Our largest data scale is 60M samples, comprising 6.6M pure-text samples from DataComp-LM (Li et al., 2024c) and 53.3M image-text samples from three sources: (a) LAION-Aesthetics (Schuhmann et al., 2022), filtered by aesthetic score ≥ 5.5 and recaptioned by InternVL3-1B (Zhu et al., 2025); (b) JourneyDB (Sun et al., 2023), recaptioned

by GPT-3.5; and (c) BLIP3o-Pretrain-Short-Caption (Chen et al., 2025a). We resize all images to 256×256 and pre-tokenize them before training.

(2) **Loss and token sequence formatting.** For pure-text samples, we adopt the standard cross-entropy loss for next-token prediction in training. For image-text samples, we initially assign 80% to text-to-image (T2I) prediction and 20% to image-to-text (I2T) prediction. As described above, 10% of the T2I samples are converted to unconditional image generation. The conditional sequence formats are:

$$\text{T2I: } \{\text{text}\}\{\text{prompt}\}\langle\text{boi}\rangle\{\textbf{image tokens}\}\langle\textbf{eoi}\rangle\langle\textbf{eos}\rangle,$$
$$\text{I2T: } \langle\text{boi}\rangle\{\text{image tokens}\}\langle\text{eoi}\rangle\{\text{prompt}\}\{\textbf{text}\}\langle\textbf{eos}\rangle.$$

Following Liquid (Wu et al., 2025a), we compute the cross-entropy loss only on the **bold** tokens (conditional next-token prediction), leaving the prompt and the conditioning modality unscored. The prompts are listed in Appendix B.2.

(3) **Hyperparameters.** We use a Warmup-Stable-Decay (WSD) schedule with a 0.03 warmup ratio and linear decay over the last 20% of steps. For the 0.6B model, we sweep the learning rate (lr) over $\{3\text{e-}5, 1\text{e-}4\}$ and the batch size (bs) over $\{512, 1024, 2048\}$, forming a 2×3 grid. Because different tasks favor different hyperparameter settings (Appendix D.2), we adopt a balanced setting of (lr $=$ 3e-5, bs $= 512$) for the main experiments and reuse it for the 1.7B and 4B models. We also experimented with larger learning rates (3e-4 for 0.6B and 1e-4 for 4B), but found that they caused unstable training with loss spikes.

Supervised finetuning (SFT) stage. We finetune the models for 2 epochs on 4.9M instruction-following samples. The mixture combines 1M LMSYS-Chat pure-text instructions (Zheng et al., 2023), 2.9M multimodal instruction and captioning samples from Mini-Gemini (Li et al., 2025), and 1M text-to-image samples (LAION-Aesthetics sampled from the continual-pretraining distribution, together with JourneyDB and BLIP3o-60K data (Chen et al., 2025a)). We use a cosine schedule with a 0.03 warmup ratio, a peak learning rate of $5\text{e-}5$, and batch size 1024.

Further details about the data used in both stages, including the subset proportions, can be found in Appendix B.1. All images used in training are publicly accessible.

Training recipe validation. We verify our training recipe by training a Qwen3-8B model with the Chameleon tokenizer (Table 2). Compared with Liquid-7B, our model reaches similar performance on GenAI-Bench, WISE, and VQA (averaged over VQAv2, GQA, TextVQA, and POPE), while trailing on MJHQ-30K. Full benchmark scores and image generation examples are provided in Appendix B.3. We view this experiment as a validation that the recipe serves as a reliable testbed for studying tokenizers, rather than as evidence of achieving superior performance.

Table 2: **Recipe validation.** Our 8B model obtains similar GenAI-Bench (Li et al., 2024a), WISE (Niu et al., 2025), and VQA (Mean) scores to Liquid-7B (Wu et al., 2025a), while trailing on MJHQ-30K (Li et al., 2024b). "Data" refers to the number of samples seen in continual pretraining.

Model	Data	GenAI↑	MJHQ-30K↓	WISE↑	VQA (Mean)↑
Liquid-7B	90M	0.72	5.47	0.41	61.40
Ours (8B)	60M	0.73	10.55	0.38	60.83

3.2 Validation Loss

Following scaling studies (Kaplan et al., 2020; Hoffmann et al., 2022; Aghajanyan et al., 2023), we report mean negative log-likelihood over supervised tokens. Let N be the evaluation set. For evaluation sequences $t \in N$, let $S(t)$ be the scored positions (bold spans above; all positions for pure text), and $T(N) = \sum_{t \in N} |S(t)|$. The model's log-likelihood on the supervised tokens is

$$l = \sum_{t \in N} \sum_{i \in S(t)} \ln p(t_i \mid t_{<i}), \tag{2}$$

We report the mean per-token loss: $\mathcal{L} = -l/T(N)$.

Note on loss interpretation. Unlike the bits-per-byte loss common in language modeling, which upper-bounds a description length of the raw data (Gao et al., 2020; Magnusson et al., 2024), our validation loss on image tokens is computed over the tokenizer's discrete *latent* codes. Because image tokenization is lossy, the model's likelihood on these codes cannot be converted into a description

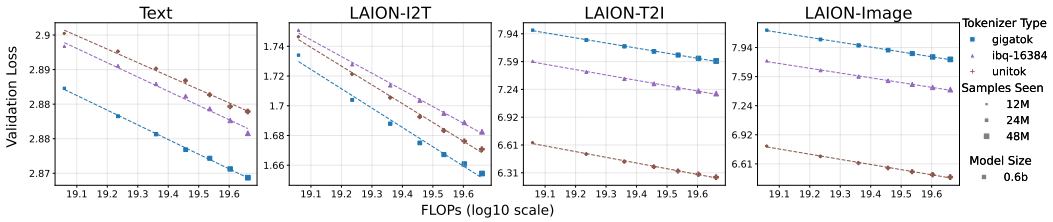

Figure 2: **Validation losses scale with data size differently across tasks** for the 0.6B model. The y-axis is on a logarithmic scale. The lines are linear fits to the dots per tokenizer.

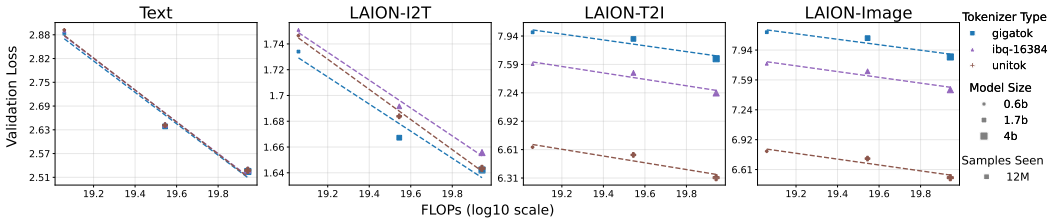

Figure 3: **Validation losses scale with model size differently across tasks** on 12M samples. The y-axis is on a logarithmic scale. The lines are linear fits to the dots per tokenizer.

length for the original pixels. We therefore interpret $\mathcal{L}$ as a measure of **latent-modeling** quality for a fixed visual language. Comparisons across image tokenizers require accounting for differences in their token spaces, as examined in Section 4.2.

Evaluation protocol. We hold out $50{,}000$ text and $50{,}000$ image-text samples from the training distribution, evaluating **text**, unconditional **image**, **T2I**, and **I2T** with the training loss masking rules. Trends across image sources are similar, so we report losses on LAION-Aesthetics (LAION-(Image, T2I, I2T)) as the proxy; losses on the other sources are given in Appendix D.1. In loss-scaling plots, the x-axis FLOPs denote the compute used in continual mixed-modal pretraining.

4 Interpreting Loss in Unified Multimodal Training

Before using pretraining loss to study tokenizers, we first examine how it should be interpreted in unified multimodal training. We begin by asking whether a single unified loss metric can characterize tokenizer behavior in downstream modeling (Section 4.1). We then examine whether these losses align with performance on image generation benchmarks and how they can be calibrated across tokenizers (Section 4.2.1). Finally, we test how they relate to post-SFT performance (Section 4.2.2).

4.1 Task-Specific Loss Scaling

Unlike language-only pretraining, our setting starts from a pretrained text model and mixes four tasks, including conditional T2I and I2T prediction. It is not obvious whether the four task-specific losses share a common scaling behavior or must be analyzed separately. We therefore first examine how each task's loss scales with data and model size, and then ask whether any single signal derived from the loss yields a tokenizer ranking that is consistent across tasks.

All task-specific losses scale with data and model size. Across three tokenizers, all four losses decrease smoothly with pretraining FLOPs, roughly following power laws under data and model scaling (Figures 2–3). For the scaling analysis, we use intermediate pre-annealing checkpoints from one run per tokenizer–model-size configuration (lr $=$ 3e-5, bs $=$ 512). We observe that annealing preserves these trends, mainly lowering text loss (Appendix D.3). Results across hyperparameters are deferred to Appendix D.2, and further continual-pretraining dynamics to Appendix C.

The scaling patterns differ across tasks. Under data scaling, text loss changes little while image, T2I, and I2T losses decrease substantially (Figure 2). Text modeling thus largely reflects the pretrained backbone, whereas image-related tasks benefit more from continual pretraining. Under model scaling, lower text loss also reflects stronger initialization and cannot be attributed to multimodal training alone. T2I closely tracks unconditional image loss, suggesting that image-token prediction accounts for much of its difficulty. I2T behaves differently from text loss despite predicting the same vocabulary, reflecting a distinct cross-modal signal.

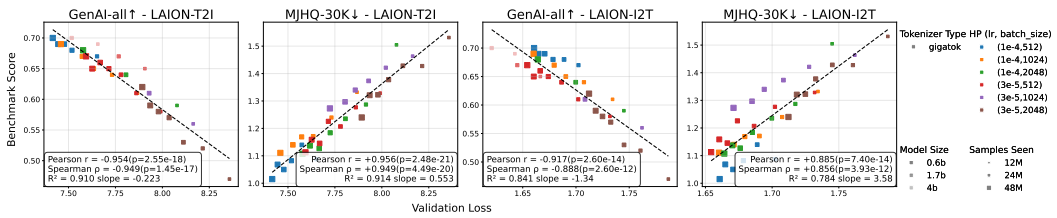

Figure 4: **Both T2I loss and I2T loss align with text-to-image generation quality when tokenizer type is controlled.** The lines are linear fits to all dots.

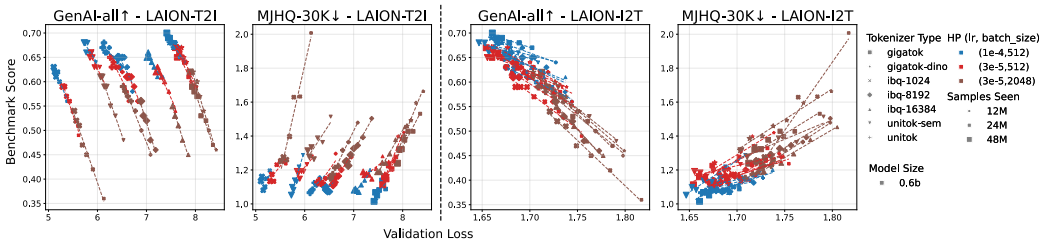

Figure 5: **T2I loss does not align with text-to-image generation quality across tokenizers** (the first two plots), **while I2T loss aligns with generation quality** (the last two plots). The lines are linear fits to the dots per tokenizer and per hyperparameter.

Tokenizer rankings differ across tasks under both data and model scaling. In Figure 2, UniTok has the highest text loss but the lowest T2I loss, while GigaTok shows the opposite pattern. Consequently, no single task loss yields a tokenizer ranking that holds across tasks, and a simple average can obscure these task-dependent differences. These results suggest trade-offs between text modeling and image-token prediction under the unified AR objective.

4.2 Loss–Performance Relation

We next study how pretraining losses relate to pre-SFT and post-SFT benchmark performance. **Benchmarks.** We evaluate the continually pretrained checkpoints on GenAI-Bench (Li et al., 2024a) and MJHQ-30K (Li et al., 2024b), scored by VQAScore (↑) and $\log_{10}$ gFID (↓), respectively. We then apply our SFT recipe to the pretraining checkpoints after annealing and re-evaluate generation performance, together with general visual understanding on VQAv2 (Goyal et al., 2017) and GQA (Hudson & Manning, 2019). Qualitative examples on GenAI-Bench and further benchmark details are in Appendix E.

4.2.1 Alignment between Losses and T2I Quality in Pretraining

We study the alignment between pretraining losses and T2I generation quality, which can be evaluated directly on the pretrained checkpoints without SFT.

Within a tokenizer, both T2I and I2T losses align with generation quality. For a fixed tokenizer, T2I loss strongly tracks GenAI-Bench VQAScore and MJHQ-30K gFID across data scales and hyperparameters (Figure 4). I2T loss also correlates, though more weakly, suggesting that it captures training progress beyond caption prediction.

Across tokenizers, T2I loss exhibits tokenizer-dependent shifts. I2T loss over the shared text vocabulary relates more consistently to generation quality (Figure 5). T2I remains a useful modeling diagnostic, but equal loss need not imply equal image quality across different visual token spaces.

Vocabulary normalization reduces the cross-tokenizer shift in the T2I loss–performance relation. We hypothesize that vocabulary size contributes to the numerical scale of T2I loss. A natural reference is $\log_2 B$, the entropy of a uniform distribution over B image tokens. We therefore define the *vocabulary-normalized* image loss as $\mathcal{L}^* = \mathcal{L}/\log_2 B$, where B denotes the image vocabulary size. Evaluated on IBQ tokenizers with $B \in \{1024, 8192, 16384\}$, $\mathcal{L}^*$ yields a more consistent T2I loss–performance relation across the IBQ variants (Figure 6). Further details are in Appendix D.6.

Residual shifts are associated with reconstruction fidelity. At a vocabulary size of 16384 and a fixed benchmark score, worse rFID corresponds to lower T2I loss (Figure 7). Reconstruction

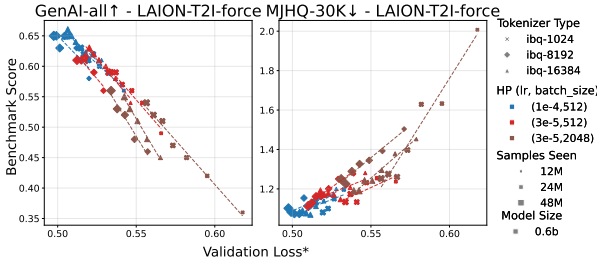

Figure 6: **After vocabulary normalization, a more consistent T2I loss–performance relation exists across IBQ tokenizers.** Quantitative results are in Table 10.

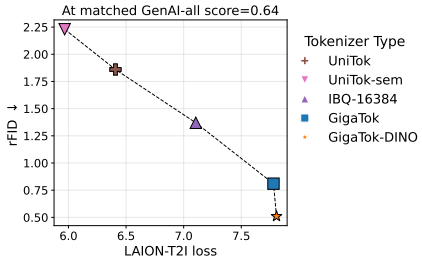

Figure 7: **The T2I loss–performance relation shift is related to differences in reconstruction fidelity.**

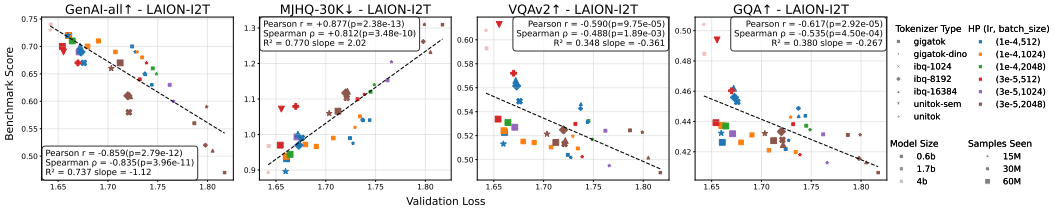

Figure 8: **I2T loss shows moderate correlation with both generation and understanding performance after SFT across tokenizers.** We use the checkpoints after annealing for loss evaluation.

fidelity may therefore help explain residual shifts and should be considered alongside T2I loss in cross-tokenizer comparisons.

4.2.2 Correlation between Losses and Post-SFT Performance

We ask whether pretraining losses remain a useful signal of benchmark performance after supervised finetuning, which introduces additional variation beyond pretraining.

Losses correlate with generation quality after SFT. The pretraining alignment of Section 4.2.1 carries over to post-SFT performance. When the tokenizer is fixed (GigaTok), both the T2I and I2T validation losses measured before SFT remain strongly correlated with post-SFT benchmark performance (Figure 20 in Appendix D.4). Across tokenizers, the I2T loss again remains the more consistent signal, staying correlated with post-SFT generation performance (Figure 8).

I2T loss is informative about general visual understanding. Across tokenizers, hyperparameters, and scales, pre-SFT I2T loss moderately correlates with post-SFT VQAv2 and GQA scores (Figure 8), indicating that it captures part of the capability that transfers to general VQA and can thus serve as a meaningful signal. The moderate correlation may partly reflect differences in task and training stage: caption prediction and post-SFT question answering differ, so caption fit is informative but not determinative. This imperfect correlation is unlikely to be explained solely by data-shuffling noise: one GigaTok configuration varies by less than 1% relative on both benchmarks across two seeds (Table 11). The relationship is less consistent for TextVQA, which additionally requires OCR (Appendix D.5).

5 What Unified Training Reveals About Image Tokenizers

We now use task-specific losses to reveal tokenizer properties that isolated metrics or single-axis evaluations miss: reconstruction fidelity can diverge from multimodal learnability, and the image token space can even affect text modeling (Section 5.2). We study these effects through three tokenizer design axes as case studies: the discriminator (Section 5.1), semantic supervision (Section 5.3), and vocabulary size (Section 5.4).

5.1 Does better reconstruction imply better unified multimodal learning?

Reconstruction-based selection assumes that higher fidelity improves downstream performance. Recent generation-only studies question this assumption (Black Forest Labs, 2025; Yao et al., 2025; Ramanujan et al., 2024; Xiong et al., 2025). We revisit this fidelity–learnability–performance relationship under unified AR training, measuring fidelity by rFID and *multimodal learnability* by task-specific validation losses.

Table 3: **Tokenizer comparison at the 0.6B model scale.** (rFID: reconstruction FID on ImageNet-1K; Text/I2T/T2I: final-step loss; GenAI and VQAv2 are post-SFT; lr=1e-4 for GigaTok family, lr=3e-5 for UniTok family, batch size=512; †: tokenizers trained by us.)

Tokenizer	Fidelity	Multimodal Learnability			Performance	
	rFID↓	Text↓	I2T↓	T2I↓	GenAI↑	VQAv2↑
GigaTok-DINO	**0.51**	2.949	**1.660**	7.437	**0.720**	51.31
GigaTok	0.81	**2.937**	1.661	**7.416**	**0.720**	**52.25**
UniTok†	**1.86**	2.883	1.670	6.264	0.670	57.21
UniTok-sem†	2.23	**2.873**	**1.655**	**5.848**	**0.690**	**61.28**

Better reconstruction does not always lead to better unified multimodal learning. Replacing PatchGAN with a DINO-based discriminator improves GigaTok's rFID from 0.81 to 0.51, but barely changes validation losses, leaves GenAI unchanged (0.720), and lowers VQAv2 from 52.25 to 51.31 (Table 3). This complements GigaTok's generation-only study (Xiong et al., 2025). Conversely, UniTok-sem has worse rFID but improves all three losses and both downstream metrics (Section 5.3). Reconstruction alone is therefore insufficient for tokenizer selection.

5.2 Does the image token space affect text modeling?

In unified AR training, text and image tokens are optimized under the same model and objective. Can the image tokenizer affect text modeling even when the text-only data, text tokenizer, and language backbone are all fixed?

Image token space affects text modeling difficulty. Under joint training (lr=3e-5, batch size=512), UniTok-sem has lower text loss than UniTok, and IBQ-1024 has lower text loss than IBQ-8192 (Figure 9, "Joint"). Thus, semantic supervision or a smaller image vocabulary can improve text modeling despite an unchanged text token space.

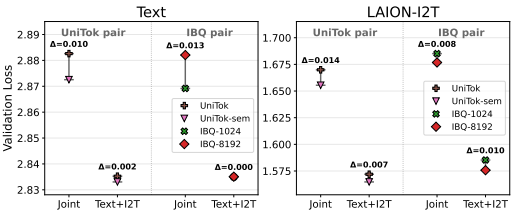

Figure 9: **Changing the image token space can reduce interference between image-token and text-token modeling in joint training.**

The text-side effect stems from image-token prediction rather than I2T. We remove image prediction by reformatting T2I and unconditional image samples into I2T order, keeping hyperparameters fixed. In this Text+I2T setting, both pairs' text-loss gaps largely disappear (Figure 9, "Text+I2T"), implicating the image-prediction objective. The I2T gaps persist, suggesting an independent effect on how informative image tokens are for captioning. A complementary I2T ablation supports this interpretation (Appendix D.7).

5.3 Semantic supervision: How does semantic alignment change multimodal learnability?

Semantic supervision improves UniTok's multimodal learnability and downstream performance despite worse reconstruction (Table 3). Does it simplify local image-token sequences ("better local visual grammar") or make tokens more informative for captioning ("better visual words")?

The I2T gain is concentrated on visually grounded object words. In Text+I2T training, we measure each tokenizer's loss reduction and compare $\Delta_{\text{UniTok-sem}} - \Delta_{\text{UniTok}}$, where larger values favor UniTok-sem. COCO object-category tokens improve much more than stopwords, general content tokens (all non-stopwords), or CSS color names (Figure 10, left).

Semantic supervision strengthens object-level image-token–word association. Validation-set pointwise mutual information (PMI) between image tokens and words, averaged within three word groups, is higher for UniTok-sem on COCO object words (Figure 10, right). This supports stronger object-level alignment.

Semantic supervision does not consistently reduce empirical image-token n-gram entropy. Following Chan et al. (2024), we measure empirical 1-, 2-, 3-, and 7-gram image-token entropy (Figure 11). UniTok-sem has slightly lower 1- and 2-gram entropy but higher 3- and 7-gram entropy, providing no consistent evidence for simpler local visual grammar under this diagnostic.

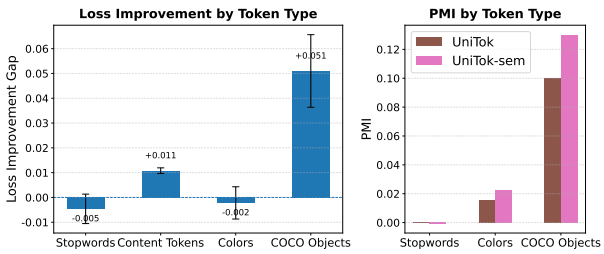

Figure 10: **Semantic supervision strengthens object-level image-token–word alignment ("better visual words").** UniTok-sem yields larger I2T loss improvements on COCO object words than on stopwords, colors, or general content tokens (left), and higher PMI between image tokens and object words (right).

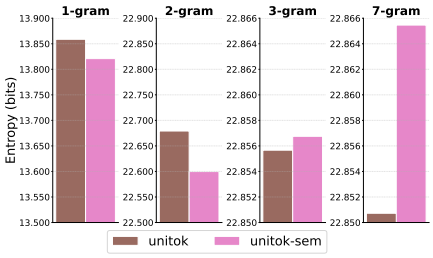

Figure 11: **Semantic supervision does not consistently reduce empirical image-token n-gram entropy.** The empirical entropy of the tokens is calculated on tokenized images of the LAION-Aesthetics validation set.

5.4 Vocabulary size: Is a larger vocabulary better for unified multimodal learning?

Larger vocabularies improve reconstruction (Shi et al., 2025), but may change modeling difficulty. We compare IBQ with $B \in \{1024, 8192, 16384\}$ using vocabulary-normalized image-side loss $\mathcal{L}^*$ (Section 4.2.1).

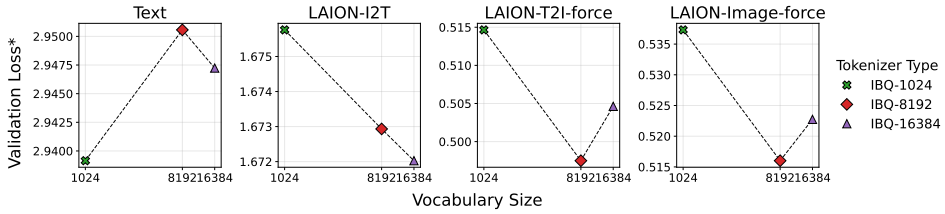

Figure 12: **Vocabulary size changes task-specific validation losses with no simple monotonic trend.** Among the three IBQ variants, IBQ-16384 achieves the best I2T loss, but IBQ-8192 attains the lowest T2I and image losses after normalization (lr=1e-4, batch size=512).

Losses vary non-monotonically with vocabulary size. IBQ-16384 has the best I2T loss, while IBQ-8192 has the lowest normalized T2I and image losses (Figure 12). Text loss follows the reverse, also non-monotonic, trend of image-side losses, again showing task-dependent preferences.

A larger vocabulary can benefit downstream performance, possibly through higher reconstruction fidelity. IBQ-16384 leads generation and understanding after SFT despite IBQ-8192's lower image-side losses (Figure 13), plausibly reflecting better reconstruction (Section 4.2.1). Whether fidelity or learnability dominates beyond $B = 16384$ remains open.

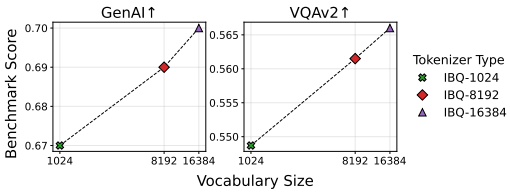

Figure 13: **The largest vocabulary brings the best downstream performance among the three IBQ variants** (lr=1e-4, batch size=512).

6 Conclusion

We study image tokenizers in the context of unified AR multimodal models through a controlled, loss-based testbed. We first examine how validation loss relates to downstream performance and how the task-specific losses should be interpreted. Building on this, we show that reconstruction fidelity can diverge from multimodal learnability and that the image token space can affect text modeling under joint training. We also revisit discriminator choice, semantic supervision, and vocabulary size to examine their effects on unified multimodal learning. Our findings suggest that unified tokenizer design should account for both reconstruction fidelity and joint modeling difficulty across tasks, rather than optimizing any single proxy metric. These observations highlight the importance of studying image tokenizers as visual languages in interaction with text during joint multimodal training.

AI USE STATEMENT

In this work, we have not used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used generative AI tools for creating scientific figures and editing the research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code for visualization and quantitative analysis was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code will be released upon acceptance, and a complete description of the training recipe and data processing steps is provided in the supplementary materials (Appendix B). All images used for training are publicly available.

REFERENCES

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023.

Black Forest Labs. FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison, 2025. URL https://bfl.ai/research/representation-comparison.

David M Chan, Rodolfo Corona, Joonyong Park, Cheol Jun Cho, Yutong Bai, and Trevor Darrell. Analyzing the language of visual tokens. *arXiv preprint arXiv:2411.05001*, 2024.

Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. ViTamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12954–12966, 2024.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. BLIP3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

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.

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, 2009.

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.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

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

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025.

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.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. GenAI-Bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024a.

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, 2024b.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. DataComp-LM: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024c.

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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.

Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. TokLIP: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025a.

Huawei Lin, Tong Geng, Zhaozhuo Xu, and Weijie Zhao. VTBench: Evaluating visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2505.13439*, 2025b.

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

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. AToken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476*, 2025.

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.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Walsh, Yanai Elazar, Kyle Lo, et al. Paloma: A benchmark for evaluating language model fit. *Advances in Neural Information Processing Systems*, 37:64338–64376, 2024.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2545–2555, 2025.

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.

Vivek Ramanujan, Kushal Tirumala, Armen Aghajanyan, Luke Zettlemoyer, and Ali Farhadi. When worse is better: Navigating the compression-generation tradeoff in visual tokenization. *arXiv preprint arXiv:2412.16326*, 2024.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16037–16046, 2025.

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

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.

Wei Song, Yuran Wang, Zijia Song, Yadong Li, Haoze Sun, Weipeng Chen, Zenan Zhou, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. DualToken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324*, 2025.

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.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

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

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond language modeling: An exploration of multimodal pretraining. *arXiv preprint arXiv:2603.03276*, 2026.

Wenxuan Wang, Fan Zhang, Yufeng Cui, Haiwen Diao, Zhuoyan Luo, Huchuan Lu, Jing Liu, and Xinlong Wang. End-to-end vision tokenizer tuning. *arXiv preprint arXiv:2505.10562*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *International Journal of Computer Vision*, 2025a.

Junfeng Wu, Dongliang Luo, Weizhi Zhao, Zhihao Xie, Yuanhao Wang, Junyi Li, Xudong Xie, Yuliang Liu, and Xiang Bai. TokBench: Evaluating your visual tokenizer before visual generation. *arXiv preprint arXiv:2505.18142*, 2025b.

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.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. GigaTok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. *arXiv preprint arXiv:2504.08736*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

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

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 VQ-GAN. *arXiv preprint arXiv:2110.04627*, 2021.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. QLIP: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 2025.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2507.08441*, 2025a.

Dian Zheng, Manyuan Zhang, Hongyu Li, Kai Zou, Hongbo Liu, Ziyu Guo, Kaituo Feng, Yexin Liu, Ying Luo, Yan Feng, et al. Architecture decoupling is not all you need for unified multimodal model. *arXiv preprint arXiv:2511.22663*, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. *Advances in Neural Information Processing Systems*, 37:12612–12635, 2024.

A Related Work

Unified multimodal understanding and generation. There have been recent advancements in building unified multimodal models, especially models for both visual understanding (captioning and VQA) and generation (text-to-image), in the hope that the two tasks will benefit each other. Various architectures have been proposed for unified models to jointly model text and images, including pure-autoregressive (AR) models (e.g., Chameleon (Team, 2024), Emu3.5 (Cui et al., 2025), Liquid (Wu et al., 2025a), Janus-Pro (Chen et al., 2025b), and LongCat-Next (Team et al., 2026)), serial AR + diffusion models (e.g., BLIP3-o (Chen et al., 2025a) and MetaQuery (Pan et al., 2025)), and hybrid AR + diffusion models (e.g., Transfusion (Zhou et al., 2024) and BAGEL (Deng et al., 2025)). Among these models, some employ a unified architecture and tokenizer for understanding and generation (e.g., Chameleon and Liquid), while others use decoupled modules, Mixture-of-Experts (MoEs), or different inference modes for the two tasks, and some adopt separate image tokenizers, illustrating the difficulty in truly unifying the two tasks in both tokenization and downstream models (Zheng et al., 2025b). We adopt a pure-AR setup to model both modalities through next-token prediction, providing a controlled setting for studying image tokenizer effects under a shared modeling objective.

Loss-based analysis of unified multimodal models. Loss-based metrics, such as Perplexity (PPL) or Bits-Per-Byte (BPB), are commonly tracked against compute to validate model efficiency: Kaplan et al. (2020) established held-out cross-entropy as a predictable scaling signal, Magnusson et al. (2024) adopt BPB because token-level perplexity is not comparable across tokenizers, and Gadre et al. (2024) fit language-model loss directly to average downstream error. Aghajanyan et al. (2023) pioneered this analysis for mixed-modal models by establishing foundational scaling laws for mixed-modal loss, demonstrating that joint cross-entropy follows predictable power laws. Chameleon uses compute-loss curves to prove the stability of its early-fusion architecture at scale (Team, 2024). More recently, Liquid (Wu et al., 2025a) utilizes these loss-vs-compute trends to demonstrate that inter-modality interference within a unified token space diminishes as model capacity increases. BAGEL (Deng et al., 2025) and Shukor et al. (2025) identify emerging reasoning properties and the advantage of early-fusion and MoEs when scaling up compute, respectively. Tong et al. (2026) use IsoFLOP analysis to uncover the scaling asymmetry between vision and language in unified models. The above works on multimodal training use loss to characterize scaling, optimization, or performance under a fixed token space; we instead ask how losses should be interpreted across image tokenizers and use losses to study image tokenizers' downstream effects.

Studying image tokenizer properties. (1) **Reconstruction.** PSNR and SSIM (Wang et al., 2004) are pixel-wise metrics for image reconstruction quality, but are sensitive to noise and poorly align with human perception. Feature-based metrics, including rFID (Heusel et al., 2017), IS (Salimans et al., 2016), and LPIPS (Zhang et al., 2018) on the ImageNet-1K validation set (Deng et al., 2009), consider semantic and distributional quality of reconstructed images. Recent tokenizer benchmarks such as VTBench (Lin et al., 2025b) and TokBench (Wu et al., 2025b) measure how tokenizers preserve text, identity, and details, which are important for downstream tasks like OCR-VQA (Mishra et al., 2019). However, better reconstruction does not necessarily translate into better downstream performance, and could conflict with better generation (Yao et al., 2025). ETT (Wang et al., 2025) shares this motivation but tunes the tokenizer end-to-end with the downstream model, whereas we keep tokenizers frozen to compare fixed visual token spaces. (2) **Generation.** Tokenizers can be compared by training with the same generative model and measuring the generation quality by gFID on a fixed set (Wu et al., 2024; Yu et al., 2024; Zhao et al., 2025; Lin et al., 2025a), but generation-only evaluation does not establish how tokenizer choice affects both generation and understanding under joint training. (3) **Semantics for understanding.** Zero-shot accuracy and linear probing accuracy on ImageNet-1K have been used for both continuous and discrete tokenizers (Radford et al., 2021; Oquab et al., 2023; Zheng et al., 2025a; Zhao et al., 2025), but higher classification accuracy does not necessarily translate into better downstream visual understanding, as observed in UniTok (Ma et al., 2025). GigaTok (Xiong et al., 2025) suggests AR probing accuracy as a better proxy to predict downstream performance when training with a larger AR model. TA-Tok (Han et al., 2025) directly measures downstream performance at varying data scales. Cambrian-1 (Tong et al., 2024) pioneered leveraging multimodal large language models as an interface for tokenizer evaluation, but is restricted to continuous tokenizers. Apart from the above, (4) **codebook usage** is often reported in earlier work, but recent tokenizers can achieve nearly 100% usage with techniques like entropy loss even when they have a large vocabulary (Zhu et al., 2024).

Table 4: Proportion of subsets in continual pretraining data. *We upsample JourneyDB from 4.2M to 6.5M.

Data Source	# Data (M)	Filtering	Recaptioning
DataComp-LM	6.6	Random Sampling	–
LAION-Aesthetics	42	Aesthetic Score (5.5)	InternVL3-1B (Zhu et al., 2025)
JourneyDB	6.5*	–	GPT3.5
BLIP3o-Pretrain-Short-Caption	4.8	–	–

Table 5: Proportion of subsets in supervised finetuning data. The LAION and JourneyDB text-to-image data is randomly sampled from the same distribution as the corresponding data in continual pretraining. *We keep all 0.2M DVQA samples without the filtering in Mini-Gemini.

Data Source	# Data (M)
Mini-Gemini (VQA)	1.7*
Mini-Gemini (Caption)	1.2
LMSYS-Chat	1.0
LAION-Aesthetics	0.86
JourneyDB	0.08
BLIP3o-60K	0.06

B Training Details

B.1 Data Proportion

We detail the data mixture in Tables 4 and 5. The SFT data are similar to those used in Liquid (Wu et al., 2025a), but we substitute the text-to-image part with public data from pretraining and use more DVQA data.

We adopt 1:8 as the text:image-text ratio in continual pretraining. This choice is influenced by the 1:2 ratio in Liquid, which uses the Chameleon tokenizer with a sequence length of 1024. We observe that it is important to keep the total number of image tokens seen during training at the same level to achieve comparable performance, so we use 1:8 as the final ratio.

B.2 Prompts

For visual generation, we uniformly sample the prompt from the same set of prompts as Liquid during pretraining: “Generate an image based on this description.”, “Create an image that captures the provided description.”, “Based on the previous text, produce a corresponding image.”, “Please illustrate the above text with a picture.”, “Translate the given description into a image.”, “Construct a visual representation of the above description.”, “Create a image that matches the text.”, “Formulate a visual expression that reflects the narrative just provided.”, “Give a visual depiction based on the above sentences.”, “Create an image using the information mentioned above as guidance.”. During loss calculation, we only use “Generate an image based on this description.” Note that for unconditional visual generation, we use the format $\langle$unconditional$\rangle\langle$boi$\rangle\{$**image tokens**$\}\langle$**eoi**$\rangle\langle$**eos**$\rangle$ without any prompt.

For captioning, we use “The caption of this image is:” as the prompt.

B.3 Model Performance

We list the model performance on visual generation and understanding benchmarks in Table 6 and provide qualitative image generation results on GenAI-Bench prompts in Figure 14. Our 8B model trained on public images with less pretraining data achieves similar performance to Liquid-7B on most benchmarks, which verifies that the framework serves as a reusable and reliable testbed for studying tokenizers’ impact on downstream unified training.

Table 6: Comparison of generation and understanding performance with Liquid-7B (Wu et al., 2025a) after SFT. "Data" refers to the number of samples seen in continual pretraining.

Model	Data	GenAI↑	MJHQ-30K↓	WISE↑	VQAv2↑	GQA↑	TextVQA↑	POPE↑	MME↑
Liquid-7B	90M	0.72	5.47	0.41	68	56.1	40.4	81.1	1107.2
Ours (8B)	60M	0.73	10.55	0.38	67.3	54.44	43.46	78.1	1038.12

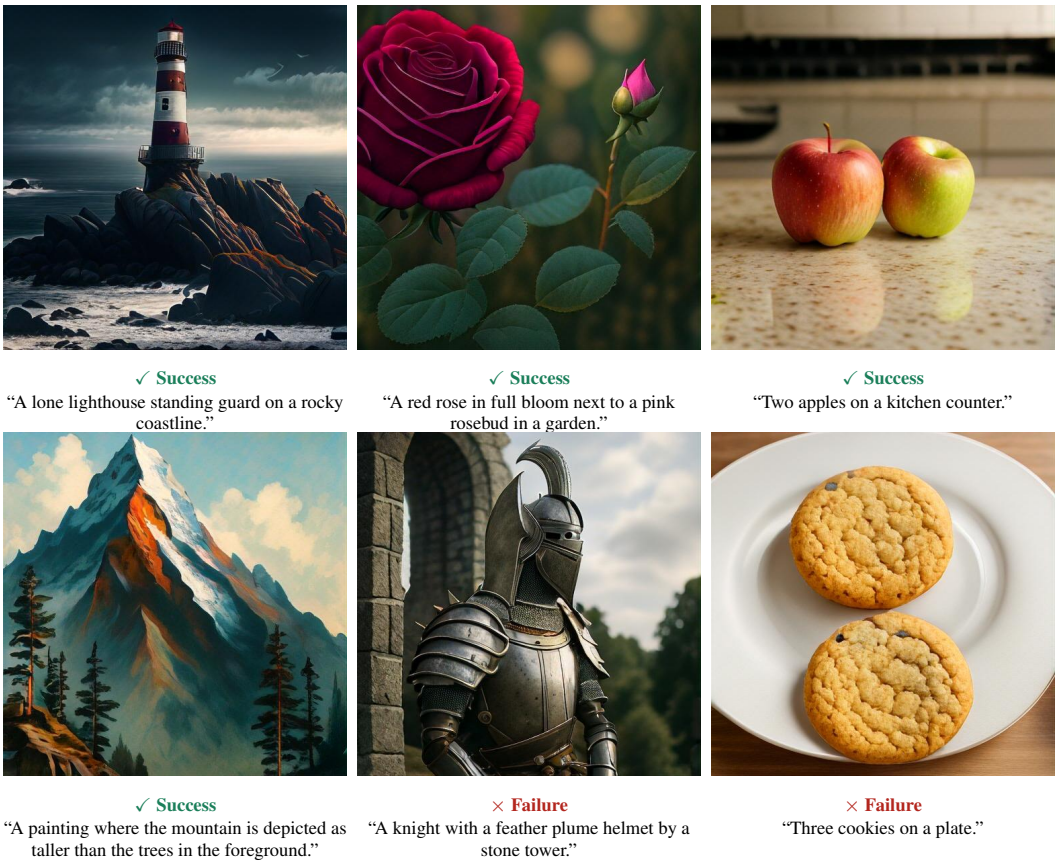

Figure 14: Qualitative text-to-image generation of our 8B model on GenAI-Bench. All images are generated by the Qwen3-8B model with the Chameleon tokenizer after SFT, i.e., the same checkpoint reported in Table 6, using the prompt shown below each panel. Green marks cases with successful instruction following and red marks failure cases.

B.4 Tokenizer Information

IBQ (Index Backpropagation Quantization) (Shi et al., 2025) modifies standard vector quantization in VQGAN to enable scalable training of large codebooks. Instead of updating only the selected code entries, IBQ backpropagates gradients to all codebook embeddings jointly with the visual encoder, which helps maintain high codebook utilization and a more consistent latent space between encoded features and code vectors. A variant of IBQ is adopted in Emu3.5 (Cui et al., 2025) with semantic loss and a vocabulary size of 131072. The authors release IBQ variants (0.1B) with vocabulary sizes of 1024, 8192, 16384, and 262144 without semantic loss. We adopt the first three since 262144 is much larger than the original vocabulary size in the base model, which might cause training difficulty.

GigaTok (Xiong et al., 2025) is proposed for autoregressive image generation and studies the effect of tokenizer scale and architectural design on discrete visual representations. It is based on VQGAN with large encoder–decoder networks and incorporates DINO-based semantic loss. It examines design choices such as asymmetric encoder–decoder scaling, different discriminator archi-

Table 7: Comparison of UniTok with and without semantic loss on rFID, top-1 zero-shot accuracy, and top-1 linear probing accuracy on ImageNet-1K. The linear probing training and evaluation follow the same practice as in the GigaTok paper.

Model	rFID↓	ZS Accuracy↑	LP Accuracy↑
UniTok	1.86	5.34	15.36
UniTok-sem	2.23	46.79	59.42

Table 8: Most tokenizers have a linear or concave curve for text but a convex curve for image-related tasks. We list the quadratic coefficient a in quadratic fitting ($y = ax^2 + bx + c$) of the 0.6B model runs in Figure 2. Positive a indicates the convexity of the curves, while negative a indicates concavity.

Tokenizer	Text	Caption	T2I	Image
GigaTok	5.46e-04	1.96e-02	6.92e-03	5.15e-03
GigaTok-DINO	1.21e-04	-7.73e-03	3.66e-03	2.54e-03
IBQ-1024	-1.63e-03	1.11e-02	2.16e-02	1.57e-02
IBQ-8192	-2.24e-03	6.94e-03	1.21e-02	1.08e-02
IBQ-16384	-2.17e-03	8.35e-03	1.15e-02	9.01e-03
UniTok	-1.59e-03	1.01e-02	1.26e-02	1.15e-02
UniTok-sem	-2.01e-03	2.19e-02	1.13e-02	1.22e-02

tectures, and alternative tokenization structures, and reports their impact on reconstruction behavior and downstream image generation. We revisit their conclusion on discriminator architectures using GigaTok-B-L with a size of 0.6B under our framework.

UniTok (Ma et al., 2025) is a unified discrete image tokenizer designed to support both image generation and visual understanding. It is also based on a VQGAN and introduces multi-codebook quantization, which partitions latent features into multiple subspaces and discretizes each with independent codebooks to increase representational capacity while maintaining stable training. The tokenizer is trained with both reconstruction and semantic loss (contrastive loss with the CLIP text encoder) so that the resulting tokens capture fine-grained visual details as well as high-level semantic information. We train a single-codebook version of UniTok with a vocabulary size of 16384 and a latent dimension of 64, and ablate the usage of semantic loss, examining its effect on downstream scaling under our framework. Following the original paper, we initialize the model with ViTamin-L/16 (Chen et al., 2024). Due to compute constraints, we train the model (0.8B) on DataComp-medium (Gadre et al., 2023) for five epochs and use 4096 as the global batch size. We adopt a smaller learning rate (lr=2e-4 for the tokenizer and 5e-6 for the discriminator) with an entropy loss weight of 0.05 and a semantic loss weight of 0 or 0.05. The rFID, zero-shot accuracy, and linear probing accuracy on ImageNet-1K of the two UniToks are shown in Table 7: UniTok-sem is stronger than UniTok in image classification but is weaker in image reconstruction.

Following the official code of UniTok, we adopt a zooming-then-cropping strategy for image preprocessing for UniTok and UniTok-sem, which differs from the preprocessing used for other tokenizers. This strategy is also applied in the pre-tokenization process in our training. Because these controlled ablations compare tokenizers within each family, this preprocessing difference does not confound those comparisons.

B.5 Compute Resources

The experiments were conducted on an internal cluster using 1, 2, 4, or 8 NVIDIA H200 GPUs with 141 GB of memory. The estimated compute for continual pretraining is listed in the result plots. The full research project also involved preliminary experiments for adjusting the training recipe and hyperparameters.

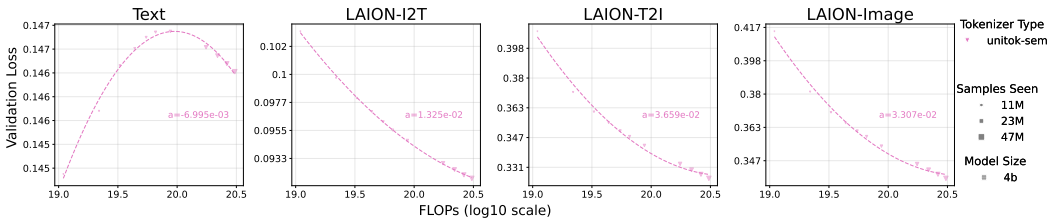

Figure 15: Quadratic fitting of UniTok-sem 4B training, including data points from an earlier stage of continual training.

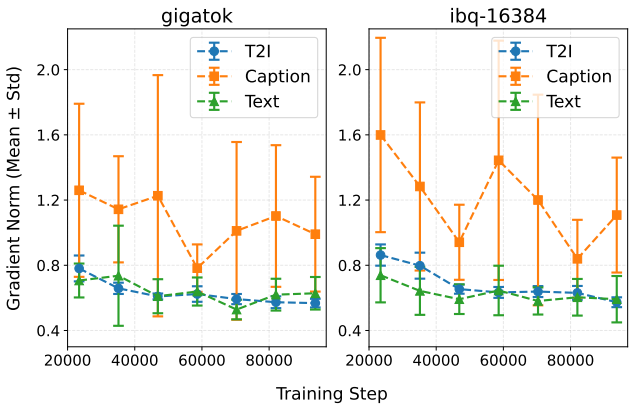

Figure 16: Gradient norms over all parameters by modality across training steps of a 0.6B model with GigaTok and IBQ-16384 (lr=3e-5, bs=512). For both runs, the T2I gradient slowly saturates at the end.

C Supplementary Results on Training Dynamics

We try quadratic fitting for the loss-data relationship in Figure 2 for all tokenizers and record the quadratic coefficients in Table 8: most data scaling curves show convexity for LAION-I2T, LAION-T2I, and LAION-Image loss, but a slight concavity for text loss. Considering that the data-scaling points correspond to checkpoints from the same runs, this implies different training dynamics across tasks. In Figure 15, we present the loss curves from training a 4B model, including earlier checkpoints, and find that their shapes align with the findings on quadratic coefficients. We hypothesize that the image-related modeling might dominate optimization in the early phase but slowly saturate later, leading to a faster decline in text loss.

We look into the gradient norm during a single run and check how the gradient norm for each modality changes over time in continual pretraining with GigaTok and IBQ-16384. In Figure 16, we find that the T2I-related gradient is larger than the text gradient at the beginning, but gradually decays throughout the training. The text gradient decays at a slower rate and reaches a similar level to the T2I gradient at the end. This supports the hypothesis about the convexity and concavity we observe in data scaling curves.

D Supplementary Loss Results

D.1 Loss over Different Image Sources

We observe very similar T2I loss and image loss trends across three data sources in Figure 17. The I2T loss trends are also similar between JourneyDB and BLIP3o-Short-Caption, but are slightly different from LAION-I2T, where (lr=3e-5, bs=512) surpasses other settings. We hypothesize that this is due to their different caption styles: recaptioned LAION-Aesthetics has longer and more detailed captions than the other two. Therefore, caption prediction on LAION-Aesthetics would rely more on the models' language ability.

D.2 Loss Scaling Results for Different Hyperparameters

To study the effect of hyperparameters on model fitting, we experiment with a 2×3 grid of (lr, bs) using GigaTok. In Figure 18, we observe that no hyperparameter setting dominates all tasks: aggressive hyperparameter settings (high lr, small bs, represented by (1e-4,512)) are favored on LAION-I2T, LAION-T2I, and LAION-Image, but lag on text loss. Mild hyperparameters represented by (3e-5, 2048) maintain the property of the base model with better text fitting. This conflict in hyperparameter preferences echoes observations on the trade-off between maintaining the pure-text ability of the base model and accommodating it for image-related tasks (Wu et al., 2025a).

D.3 Annealed Loss Results

We plot the loss-FLOPs results after annealing along with the non-annealed results in Figure 19. On all tasks, the relative ranking among hyperparameters is barely changed after annealing compared with results before annealing in Figure 18. One notable exception is that lr=1e-4 has a greater reduction in loss during annealing: on LAION-I2T, (lr=3e-5, bs=512) reaches the lowest loss at the end of the constant learning rate stage, but is outperformed by (lr=1e-4, bs=512) after annealing. We observe that annealing mainly reduces the text loss but does not affect the image-related losses significantly. One hypothesis is that the loss on image tokens might lead to noisier or smaller gradients during the annealing stage than the loss on the text tokens.

D.4 Loss and Post-SFT Generation Performance for a Fixed Tokenizer

Figure 20 shows the relationship between pre-SFT validation losses and post-SFT generation performance for GigaTok.

D.5 Loss-VQA Performance on TextVQA

We study the relationship between pretraining loss and a more specific VQA task: TextVQA. In Figure 21, I2T loss is negatively correlated with post-SFT TextVQA performance for 0.6B models, but shows the opposite trend when we scale up the model size to 4B. As a side finding, we observe that optimal hyperparameters vary for different benchmarks: aggressive hyperparameters win on VQAv2 and GQA, but mild hyperparameters lead to better TextVQA performance despite higher

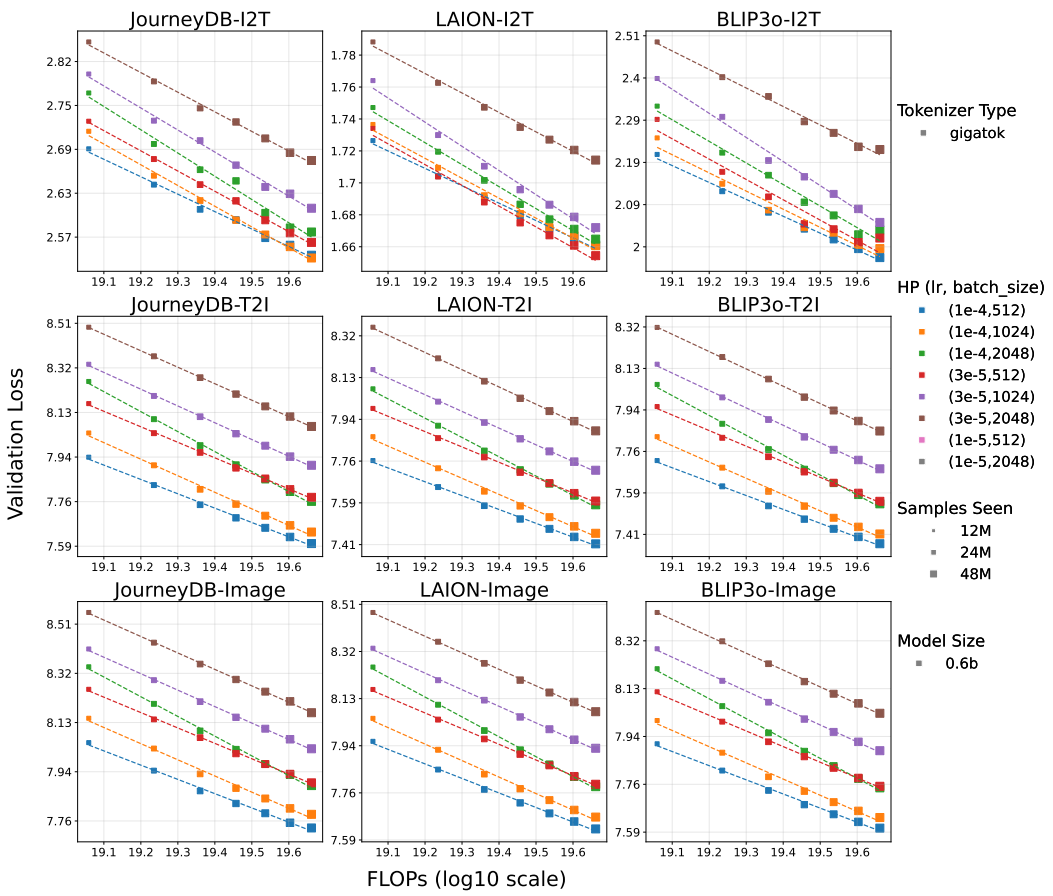

Figure 17: Loss scales with data similarly for image-text data from different sources.

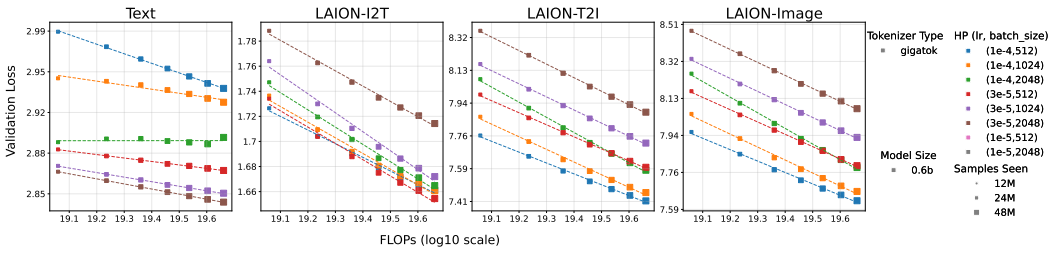

Figure 18: Loss scales with data for the 0.6B model with GigaTok differently under six hyperparameter settings. Text loss prefers a small learning rate and a large batch size, while image-related tasks favor the opposite settings.

I2T loss. These results suggest that post-SFT understanding performance sometimes depends on additional capabilities beyond pretraining caption fit, such as OCR. On the other hand, treating the average score on several VQA benchmarks as a measure of visual understanding performance might omit their distinct properties and could be highly sensitive to benchmark choices.

D.6 Details on Loss Normalization

For LAION-T2I-force and LAION-Image-force, we first set the output logits for non-image tokens to -inf between ⟨boi⟩ and ⟨eoi⟩. This ensures that the model can only choose from the image tokens. Then, we normalize the loss (log perplexity) by log vocabulary size. Empirically, we find that the effect of the first step (forcing image output) is negligible for the loss values, indicating that the

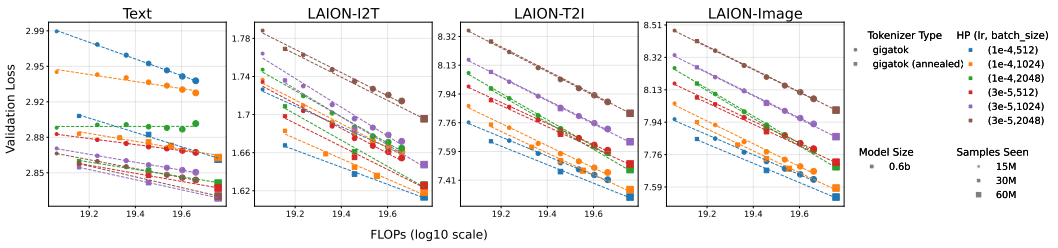

Figure 19: Loss scales with data for annealed and non-annealed checkpoints.

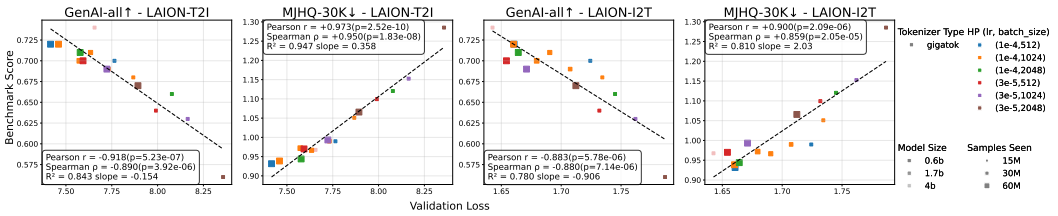

Figure 20: **Both I2T and T2I loss show strong correlation with post-SFT generation performance when tokenizer type is controlled.** We use the checkpoints after annealing for loss evaluation.

model has learned to output image tokens between ⟨boi⟩ and ⟨eoi⟩ on the validation set even at the first checkpoint plotted in the results (trained with 12M data).

Note that we do not apply the same normalization to the text loss and I2T loss even though the total vocabulary sizes are not identical for the three IBQ variants. The reason is that we also find the first step (forcing text output) has a negligible effect on the loss values, which shows that the model only chooses from text tokens during text and I2T evaluation on the validation set. Hence, we do not normalize the losses by the total vocabulary size.

Comparison with empirical code entropy. Normalizing by $\log_2 B$ assumes the maximal entropy of the image vocabulary, whereas real codebooks are used unevenly. To test whether $\log_2 B$ is an adequate stand-in for the actual code distribution, we compare it with the empirical unigram entropy H_1 measured on the validation image tokens of all seven tokenizers (Table 9). H_1 lies within 1.3% of $\log_2 B$ for every tokenizer. Uneven code usage does occur: GigaTok's rank-frequency distribution is visibly skewed, with head codes used roughly 40 times as often as tail codes. This skew is nevertheless confined to a small fraction of code ranks and carries little probability mass, so GigaTok's entropy stays close to $\log_2 B$. The larger deficits of IBQ-16384 and UniTok-sem instead reflect lower utilization across a broader portion of the codebook.

Substituting H_1 for $\log_2 B$. We then relate the normalized T2I loss to GenAI-all at the 0.6B scale for the three IBQ variants under both normalizations (Table 10). The two choices give nearly identical results: both recover a strong loss–performance relation, whereas the unnormalized loss explains essentially none of the benchmark variance and tokenizer identity accounts for most of it.

Scope of the normalization. This normalization corrects the loss scale associated with vocabulary size; it does not calibrate losses across tokenizer architectures at a fixed B, which is consistent with the cross-tokenizer clustering in Figure 5. Our experiments also do not cover tokenizers with severe or complete codebook collapse; in such a regime the nominal vocabulary size would no longer represent the effective prediction space, and $\log_2 B$ normalization may not be applicable.

D.7 Ablation on I2T Objective

To verify the effect of the I2T objective in joint training, we ablate it by reformatting the I2T samples into T2I order (with a drop rate of 10%, which enables CFG) and train with only Text+T2I data with the same hyperparameter settings. This experimental setting resembles the ablation in Section 5.2 for studying the effect of the T2I objective.

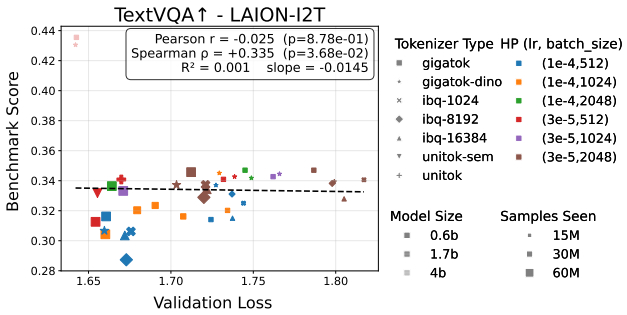

Figure 21: I2T loss has a noisy and inconsistent correlation with TextVQA performance.

Table 9: Maximal entropy $\log_2 B$ versus empirical unigram entropy H_1 of the image tokens on the validation set. H_1 is within 1.3% of $\log_2 B$ for all seven tokenizers.

Tokenizer	$\log_2 B$	H_1	$H_1 / \log_2 B$
GigaTok	14.000	13.953	0.997
GigaTok-DINO	14.000	13.949	0.996
IBQ-1024	10.000	9.977	0.998
IBQ-8192	13.000	12.947	0.996
IBQ-16384	14.000	13.841	0.989
UniTok	14.000	13.859	0.990
UniTok-sem	14.000	13.821	0.987

In Figure 22, we observe that the gap between tokenizers on text and T2I is not reduced by ablating I2T data. Notably, the text loss increases for both UniTok and UniTok-sem when the I2T objective is ablated. This suggests that the I2T objective complements, rather than competes with, text modeling during joint training. For comparison, the T2I loss slightly decreases when there is no I2T task.

E Supplementary Benchmark Information and Results

(1) Text-to-image generation. Following Liquid (Wu et al., 2025a), we report (1) VQAScore on GenAI-Bench (Li et al., 2024a), calculated by asking CLIP-FlanT5-XXL whether the image aligns with the text and recording the probability of answering "Yes" (hence ranging from 0 to 1), and (2) gFID on MJHQ-30K (Li et al., 2024b).During inference, we use unfiltered softmax sampling and CFG scale=7.0. We adopt a unified template "{text} Generate an image based on this description.⟨boi⟩" for evaluation. Note that we force the model output to be an image token by setting the probability of text tokens to zero during sampling. Without this operation, we find that an under-optimized unified model could output mixed text tokens and image tokens, which even holds for larger public models like Liquid-7B. Some prompts and generated results are in Figure 23. The images are generated by the Qwen3-0.6B model trained with lr=1e-4 and bs=512 on 60M samples. The qualitative results align with the reported benchmark scores, as GigaTok and GigaTok-DINO produce images with the best quality and alignment, followed by UniTok and UniTok-sem with worse quality but good alignment.

(2) Visual understanding. We use VQAv2 (Goyal et al., 2017) and GQA (Hudson & Manning, 2019) as two general VQA benchmarks. TextVQA (Singh et al., 2019) is included in Appendix D.5 as a specific VQA benchmark. We also test the models on POPE (Li et al., 2023) and MME (Yin et al., 2024) following Liquid, but find large variance among runs with different shuffling seeds (see Table 11 and the analysis below). Therefore, we exclude them from our analysis.

Benchmark sensitivity to random seeds. All central experiments in this work use a single data-shuffling seed (42), which controls the ordering of training data while the language-model initialization is unchanged. To quantify how much of the benchmark spread this ordering accounts for, we additionally ran SFT on one GigaTok configuration (0.6B model, lr=1e-4, bs=1024) with data-shuffling seed 37 and compared the two runs (Table 11). VQAv2, GQA, GenAI-all, and MJHQ-30K agree to within 1.9% relative difference, whereas POPE changes by 7.07% and MME-P by 4.98%. This quantitatively motivates excluding POPE and MME-P from the loss-correlation analysis: their data-order sensitivity is substantially larger than that of the benchmarks we do analyze. We note that

Table 10: Relating normalized T2I loss to GenAI-all across the three IBQ variants (0.6B models). Normalizing by the empirical entropy H_1 is nearly equivalent to normalizing by $\log_2 B$, while the unnormalized loss leaves the variance to be explained by tokenizer identity.

Normalization	Pearson r	R^2 from loss	Additional R^2 from tokenizer identity
None	$+0.016$	0.000	0.972
Divide by $\log_2 B$	-0.957	0.915	0.048
Divide by H_1	-0.953	0.909	0.055

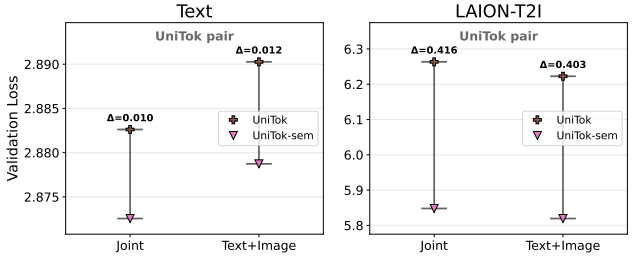

Figure 22: In the joint modeling setting, UniTok-sem has lower text loss than UniTok. In Text+T2I training ablating the I2T task, the gap is not reduced, suggesting that the interference does not come from the I2T task. As a side finding, the I2T task complements text modeling in joint training for both tokenizers.

this is a two-seed comparison for a single GigaTok configuration rather than a multi-seed replication of every tokenizer comparison due to high computational cost.

F LIMITATIONS

First, our study is limited to seven discrete image tokenizers with a fixed token length and single-codebook designs to maintain a controlled experimental setting. For example, on the sequence length (K) axis, we observe that mixed-modal training is sensitive to the number of image tokens seen. Since it is mathematically impossible to simultaneously hold both the number of image tokens seen and the number of images seen constant when K varies, such a study would introduce confounding variables. We leave the study of tokenizers with sub-codebooks (Lu et al., 2025), larger input resolutions (Team, 2024), and different sequence lengths (Yu et al., 2024) to future research.

Second, we use pretraining validation loss as a lens for studying joint modeling behavior, not as a replacement for downstream evaluation after SFT or for tokenizer selection. Although I2T loss provides a useful signal, it is less predictive for specialized tasks such as TextVQA.

Third, due to the high computational cost, we do not conduct a full scaling-law study of unified multimodal AR training that starts from scratch with more model scales or derive compute-efficient frontiers. Future work could derive and compare image tokenizers' effects on the compute-efficient frontier of multimodal AR models.

Fourth, analyses on the validation set, including PMI and n-gram entropy, are diagnostic rather than mechanistic, and high-order empirical entropy can suffer from finite-sample bias.

Fifth, our empirical findings are established for Qwen3-based pure-AR unified models. Holding the backbone family fixed is itself a control in our study, since it lets us vary the visual token space while keeping the downstream architecture and training recipe consistent. We therefore do not claim that every loss relationship or tokenizer ranking is unchanged under a different backbone family; validating them would require matched retraining and evaluation across tokenizers under that backbone. The analysis framework itself can be applied to another language-model family, and we view such cross-family validation as an important direction for future work.

Finally, our downstream evaluation covers a limited set of generation and VQA benchmarks; broader task, safety, and human-preference evaluations are needed to fully evaluate the models. In particular, our understanding evaluation covers general VQA (VQAv2 and GQA) and the OCR-oriented TextVQA, but includes no dedicated region-level grounding or fine-grained spatial-reasoning bench-

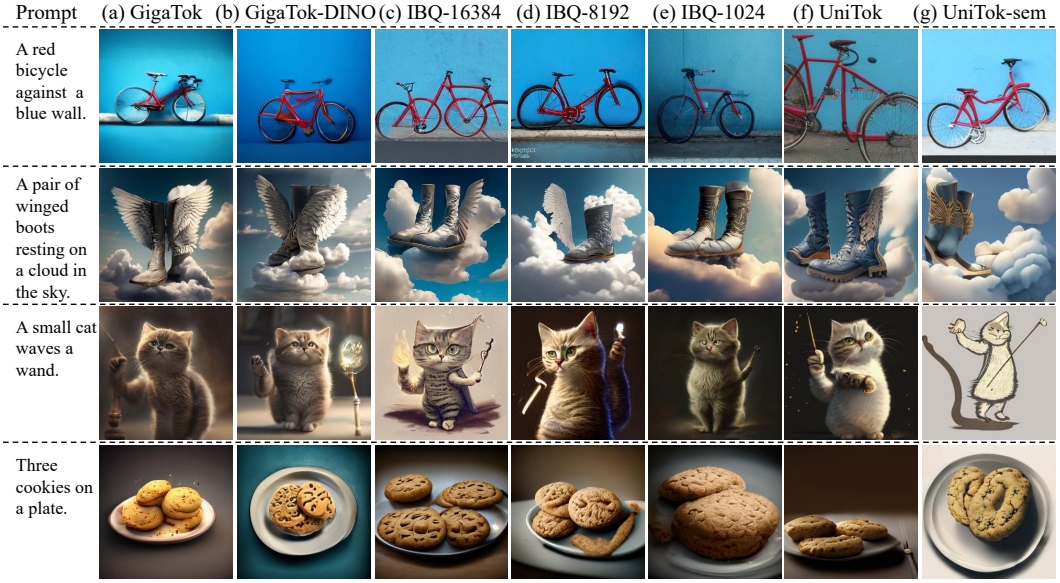

Figure 23: Qualitative comparison of generation quality on GenAI-Bench.

Table 11: Effect of the data-shuffling seed on post-SFT benchmark scores for one GigaTok configuration (0.6B, lr=1e-4, bs=1024). Relative difference is the absolute difference divided by the mean of the two runs. POPE and MME-P are markedly more sensitive to random seeds than the other benchmarks.

Benchmark	Seed 42	Seed 37	Abs. diff.	Rel. diff.
VQAv2↑	51.43	51.47	0.04	0.08%
GQA↑	43.11	42.71	0.40	0.93%
POPE↑	61.40	65.90	4.50	7.07%
MME-P↑	825.56	785.48	40.08	4.98%
GenAI-all↑	0.710	0.700	0.010	1.42%
MJHQ-30K gFID↓	9.2558	9.0812	0.1746	1.90%

mark. We therefore do not claim that our findings extend to these specialized visual capabilities, and evaluating how the image token space affects them remains future work.

G Third-Party Attribution Notices

Portions of this work use third-party software released under the MIT License: we train UniTok and UniTok-sem variants using the official UniTok code (`https://github.com/FoundationVision/UniTok`), and we use the GigaTok and GigaTok-DINO tokenizers (`https://github.com/SilentView/GigaTok`). The copyright and permission notices below are reproduced as required by that license.

> Copyright (c) 2024 FoundationVision
>
> Copyright (c) 2025 GigaTok: Scaling Visual Tokenizers to 3 Billion Parameters for Autoregressive Image Generation Authors
>
> Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

H BROADER IMPACTS

This work studies image tokenizers in the context of unified pure-AR multimodal models. By showing how tokenizer design affects modeling, including cross-modal modeling and text-side behavior, our analysis may help researchers design and diagnose multimodal training beyond reconstruction or single-axis evaluation. Improved tokenizer analysis could contribute to stronger generation and understanding models, which may support creative, educational, and accessibility applications.

However, stronger multimodal models also raise risks such as biased outputs, misleading synthetic content, and misuse. Our experiments use publicly available images and do not introduce a deployed system, but future work using these insights should be paired with standard safety, bias, privacy, and responsible-release evaluations.