Novelty Assessment Report

Paper: Latent Denoising Makes Good Visual Tokenizers
PDF URL: https://openreview.net/pdf?id=1jBsi98fVe
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Despite their fundamental role, it remains unclear what properties could make tokenizers more effective for generative modeling. We observe that modern generative models share a conceptually similar training objective---reconstructing clean signals from corrupted inputs, such as signals degraded by Gaussian noise or masking---a process we term denoising. Motivated by this insight, we propose aligning tokenizer embeddings directly with the downstream denoising objective, encouraging latent embeddings that remain reconstructable even under significant corruption. To achieve this, we introduce the Latent Denoising Tokenizer (l-DeTok), a simple yet highly effective tokenizer trained to reconstruct clean images from latent embeddings corrupted via interpolative noise or random masking. Extensive experiments on class-conditioned (ImageNet 256x256 and 512x512) and text-conditioned (MSCOCO) image generation benchmarks demonstrate that our l-DeTok consistently improves generation quality across six representative generative models compared to prior tokenizers. Our findings highlight denoising as a fundamental design principle for tokenizer development, and we hope it could motivate new perspectives for future tokenizer design. Our code and models will be publicly available.

Core Task Landscape

This paper addresses: **Visual Tokenizer Design for Generative Modeling**
A total of **50 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Tokenization Architecture and Representation Space**
- **Training Objectives and Optimization Strategies**
- **Unified Tokenizers for Understanding and Generation**
- **Tokenizers for Autoregressive Generation Models**
- **Multimodal and Cross-Domain Tokenization**
- **Domain-Specific and Application-Oriented Tokenization**

Complete Taxonomy Tree

- Visual Tokenizer Design for Generative Modeling Survey Taxonomy
- Tokenization Architecture and Representation Space
  - Discrete Token-Based Architectures
  - Standard VQ-Based Tokenizers (3 papers)
    - [14] Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors (Gafni, 2022) View paper
    - [26] CogView: Mastering Text-to-Image Generation via Transformers (Ding Ming, 2021) View paper
    - [31] MaskGIT: Masked Generative Image Transformer (Huiwen Chang, 2022) View paper
  - Large-Scale Codebook Tokenizers (2 papers)
    - [12] Open-magvit2: An open-source project toward democratizing auto-regressive visual generation (Luo, 2024) View paper
    - [15] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation (Yu, 2023) View paper
  - Factorized and Hierarchical Quantization (3 papers)
    - [19] Factorized visual tokenization and generation (Bai, 2024) View paper
    - [25] WeTok: Powerful Discrete Tokenization for High-Fidelity Visual Reconstruction (Zhuang Shaobin, 2025) View paper
    - [32] ImageFolder: Autoregressive Image Generation with Folded Tokens (Li Xiang, 2024) View paper
  - Lookup-Free and Alternative Quantization Schemes (3 papers)
    - [22] QLIP: Text-Aligned Visual Tokenization Unifies Auto-Regressive Multimodal Understanding and Generation (Zhao Yue, 2025) View paper
    - [33] Xq-gan: An open-source image tokenization framework for autoregressive generation (Li Xiang, 2024) View paper
    - [48] BiGR: Harnessing Binary Latent Codes for Image Generation and Improved Visual Representation Capabilities (Hao, 2024) View paper
  - Continuous Latent Space Tokenizers (3 papers)
  - [9] Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens (Fan Lijie, 2024) View paper
  - [18] V2Flow: Unifying Visual Tokenization and Large Language Model Vocabularies for Autoregressive Image Generation (Zhang Gui-Wei, 2025) View paper
  - [35] Ming-UniVision: Joint Image Understanding and Generation with a Unified Continuous Tokenizer (Huang, 2025) View paper
  - Spatial-Temporal and Multi-Resolution Architectures (4 papers)
  - [10] Scalable Visual Tokenizers for Generative Modeling (Yin-bo, 2025) View paper
  - [13] Atoken: A unified tokenizer for vision (Lu, 2025) View paper
  - [36] LARP: Tokenizing Videos with a Learned Autoregressive Generative Prior (Wang Hanyu, 2024) View paper

- [38] OmniTokenizer: A Joint Image-Video Tokenizer for Visual Generation (Wang Jun-ke, 2024) View paper
- Holistic and Query-Based Tokenization (1 papers)
- [39] Holistic Tokenizer for Autoregressive Image Generation (A Zheng, 2025) View paper
- Spectral and Transform-Domain Tokenization (1 papers)
- [46] Spectral image tokenizer (Esteves, 2025) View paper

- Training Objectives and Optimization Strategies
  - Reconstruction-Focused Training (2 papers)
  - [40] Flow to the Mode: Mode-Seeking Diffusion Autoencoders for State-of-the-Art Image Tokenization (Sargent, 2025) View paper
  - [42] Learnings from scaling visual tokenizers for reconstruction and generation (Hansen-Estruch, 2025) View paper
  - Denoising-Based Training Objectives ★ (2 papers)
  - [0] Latent Denoising Makes Good Visual Tokenizers (Anon et al., 2026) View paper
  - [27] Layton: Latent Consistency Tokenizer for 1024-pixel Image Reconstruction and Generation by 256 Tokens (Xie Qing-song, 2025) View paper
  - Post-Training and Generation-Aware Refinement (2 papers)
  - [5] Image tokenizer needs post-training (Qiu Kai, 2025) View paper
  - [29] Stabilize the latent space for image autoregressive modeling: A unified perspective (Yongxin Zhu, 2024) View paper
  - Latent Space Stabilization Techniques (1 papers)
  - [41] ResiComp: Loss-Resilient Image Compression via Dual-Functional Masked Visual Token Modeling (Sixian Wang, 2025) View paper
- Unified Tokenizers for Understanding and Generation
  - Dual-Objective Unified Tokenizers (3 papers)
  - [2] UniTok: A Unified Tokenizer for Visual Generation and Understanding (Ma, 2025) View paper
  - [4] Tokenflow: Unified image tokenizer for multimodal understanding and generation (Liao Qu, 2025) View paper
  - [16] Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation (Chen Zisheng, 2025) View paper
  - Foundation Model Alignment for Tokenization (2 papers)
  - [23] Vision foundation models as effective visual tokenizers for autoregressive image generation (Zheng, 2025) View paper
  - [24] Aligning Visual Foundation Encoders to Tokenizers for Diffusion Models (Chen Bowei, 2025) View paper
- Tokenizers for Autoregressive Generation Models
  - Masked Generative Tokenization (3 papers)
  - [3] Muse: Text-To-Image Generation via Masked Generative Transformers (Chang, 2023) View paper
  - [47] MAGE: MAsked Generative Encoder to Unify Representation Learning and Image Synthesis (Tianhong Li, 2022) View paper
  - [49] Resurrect mask autoregressive modeling for efficient and scalable image generation (Xin Yi, 2025) View paper
  - Order-Agnostic and Flexible Generation Tokenizers (1 papers)
  - [11] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders (Ziqi Pang, 2024) View paper
  - Reinforcement Learning Enhanced Tokenization (1 papers)
  - [17] X-Omni: Reinforcement Learning Makes Discrete Autoregressive Image Generative Models Great Again (Geng, 2025) View paper
  - Scaling and Benchmarking Autoregressive Tokenizers (2 papers)
  - [7] Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation (Xiong Tian-wei, 2025) View paper
  - [37] VTBench: Evaluating Visual Tokenizers for Autoregressive Image Generation (Lin, 2025) View paper
- Multimodal and Cross-Domain Tokenization
  - Vision-Language Unified Tokenizers (2 papers)
  - [8] VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation (Wu Yecheng, 2024) View paper
  - [20] ILLUME+: Illuminating Unified MLLM with Dual Visual Tokenization and Diffusion Refinement (Huang, 2025) View paper
  - Video-Language Tokenization (1 papers)
  - [50] Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization (Jin Yang, 2024) View paper
  - Prompt Tuning and Transfer Learning for Tokenizers (1 papers)
  - [6] Visual Prompt Tuning for Generative Transfer Learning (Kihyuk Sohn, 2022) View paper
- Domain-Specific and Application-Oriented Tokenization
  - Text-to-Image Generation Tokenizers (1 papers)
  - [45] Empowering Backbone Models for Visual Text Generation with Input Granularity Control and Glyph-Aware Training (Li Wenbo, 2024) View paper
  - Style Transfer and Artistic Generation Tokenizers (1 papers)
  - [44] Stylerdalle: Language-guided style transfer using a vector-quantized tokenizer of a large-scale generative model (Zipeng Xu, 2023) View paper
  - Long-Form Video and Narrative Generation (1 papers)
  - [43] MovieDreamer: Hierarchical Generation for Coherent Long Visual Sequence (Liu, 2024) View paper
  - Specialized Domain Tokenization (2 papers)
  - [21] OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving (Julong Wei, 2024) View paper
  - [34] Language of Stains: Tokenization Enhances Multiplex Immunofluorescence and Histology Image Synthesis (Zachary Sims, 2025) View paper
  - Retrieval and Recommendation Tokenization (2 papers)
  - [1] Learning to tokenize for generative retrieval (Sun Wei-wei, 2023) View paper
  - [28] A Simple Contrastive Framework Of Item Tokenization For Generative Recommendation (Zhai, 2025) View paper
  - Vision-Language-Action Tokenization (1 papers)
  - [30] A Survey on Vision-Language-Action Models: An Action Tokenization Perspective (Zhong, 2025) View paper

Narrative

Core task: visual tokenizer design for generative modeling. The field organizes around several complementary dimensions. One branch explores tokenization architecture and representation space, examining how discrete codes or continuous embeddings best capture visual structure. Another focuses on training objectives and optimization strategies, including reconstruction losses, adversarial training, and denoising-based formulations that guide tokenizers toward semantically meaningful representations. A third branch investigates

unified tokenizers that bridge understanding and generation tasks, while a fourth targets tokenizers optimized specifically for autoregressive generation models. Additional branches address multimodal and cross-domain tokenization—extending visual codes to language, audio, or video—and domain-specific applications such as medical imaging or robotics. Representative works like Muse[3] and MaskGIT[31] illustrate how different training regimes shape tokenizer behavior, while efforts such as VILA-U[8] and OmniTokenizer[38] demonstrate the push toward unified multimodal representations.

Within the training objectives branch, a particularly active line of work explores denoising-based formulations that encourage tokenizers to learn robust, noise-invariant features. Latent Denoising Tokenizers[0] exemplifies this approach by incorporating denoising objectives directly into the tokenization process, contrasting with purely reconstruction-driven methods. Nearby, Layton[27] also emphasizes denoising mechanisms but may differ in architectural choices or the balance between reconstruction fidelity and semantic abstraction. Meanwhile, post-training refinement strategies such as Tokenizer Post-training[5] adjust pretrained tokenizers to better align with downstream generative models, highlighting an ongoing tension between end-to-end joint training and modular design. The original paper sits squarely in this denoising-focused cluster, contributing to the broader question of how noise-aware objectives can yield tokenizers that generalize across diverse generative architectures and data distributions.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Layton: Latent Consistency Tokenizer for 1024-pixel Image Reconstruction and Generation by 256 Tokens

Authors: Xie Qing-song, Zhang Zhao, Qingsong Xie, Huang Zhe, Zhao Zhang, et al. (13 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Image tokenization has significantly advanced visual generation and multimodal modeling, particularly when paired with autoregressive models. However, current methods face challenges in balancing efficiency and fidelity: high-resolution image reconstruction either requires an excessive number of tokens or compromises critical details through token reduction. To resolve this, we propose Latent Consistency Tokenizer (Layton) that bridges discrete visual tokens with the compact latent space of pre-...

Relationship Analysis

Both papers belong to the denoising-based training objectives category, training tokenizers to reconstruct clean signals from corrupted latent embeddings. They overlap in using denoising processes during tokenizer training to improve downstream generative modeling, with both applying noise corruption to latent representations. However, the original paper (l-DeTok) focuses on interpolative latent noise and random masking to align with diverse generative models (both AR and non-AR), while Layton specifically integrates with pre-trained Latent Diffusion Models using a latent consistency decoder that reduces multi-step sampling to 1-2 steps for extreme compression (256 tokens for 1024x1024 images).

Contributions Analysis

Overall novelty summary. The paper proposes l-DeTok, a tokenizer trained to reconstruct clean images from latent embeddings corrupted via interpolative noise or random masking, positioning denoising as a core design principle. It resides in the 'Denoising-Based Training Objectives' leaf, which contains only two papers including this one. This is a notably sparse research direction within the broader taxonomy of 50 papers across 27 leaf nodes, suggesting that explicit denoising-based tokenizer training remains relatively underexplored compared to reconstruction-focused or post-training refinement approaches.

The taxonomy reveals that l-DeTok's immediate neighbors include reconstruction-focused training methods and post-training refinement strategies, both of which emphasize alignment with downstream objectives but differ in mechanism. Nearby branches address latent space stabilization and generation-aware refinement, indicating that the field is actively exploring how to bridge tokenizer learning and generative model requirements. The denoising-based leaf sits within a larger branch on training objectives, distinct from architectural innovations like factorized quantization or continuous latent spaces, clarifying that l-DeTok's novelty centers on the training regime rather than representational structure.

Among 30 candidates examined, the first contribution—l-DeTok as a method—shows one refutable candidate out of 10 examined, suggesting some prior work addresses denoising-based tokenizer training. The second contribution, framing denoising as a unifying principle, examined 10 candidates with none clearly refuting it, indicating this conceptual framing may be less directly anticipated. The third contribution, comprehensive empirical validation across six generative models, also examined 10 candidates with no refutations, suggesting the breadth of experimental coverage is relatively distinctive within the limited search scope.

Given the sparse population of the denoising-based training leaf and the limited search scale, the work appears to occupy a relatively novel position in explicitly aligning tokenizer objectives with downstream denoising processes. However, the presence of one refutable candidate for the core method indicates that the technical approach may have partial precedent. The analysis reflects top-30 semantic matches and does not constitute an exhaustive survey of all tokenizer training strategies or denoising formulations in the broader literature.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Latent Denoising Tokenizer (l-DeTok)

Description: The authors propose a tokenizer training method that aligns latent embeddings with downstream generative model objectives by reconstructing clean images from corrupted latent representations. This is achieved through interpolative noise injection and optional random masking during training, encouraging robust and easily reconstructable embeddings.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Brain-Semantoks: Learning Semantic Tokens of Brain Dynamics with a Self-Distilled Foundation Model

URL: View paper

Brief Assessment

Brain-Semantoks[63] focuses on fMRI time series analysis using semantic tokenization of brain networks with self-distillation objectives, not on image generation tokenizers trained with noise injection for reconstructing clean images from corrupted latent embeddings.

---

2. Layton: Latent Consistency Tokenizer for 1024-pixel Image Reconstruction and Generation by 256 Tokens

URL: View paper

Brief Assessment

Layton[27] focuses on bridging discrete visual tokens with latent diffusion models for efficient compression, not on training tokenizers to reconstruct from corrupted latent embeddings via interpolative noise injection as in the original paper.

---

3. LacTok: Latent Consistency Tokenizer for High-resolution Image Reconstruction and Generation by 256 Tokens

URL: View paper

Brief Assessment

LacTok[64] focuses on bridging discrete visual tokens with pretrained latent diffusion models for high-resolution image reconstruction, not on training tokenizers to reconstruct from corrupted latent embeddings using interpolative noise injection and masking as proposed in the original paper.

---

4. X-lrm: X-ray large reconstruction model for extremely sparse-view computed tomography recovery in one second

URL: View paper

Brief Assessment

X-lrm[57] focuses on CT reconstruction from X-ray projections using transformer-based encoders and triplane representations for medical imaging. The paper does not address tokenizer training for generative models or methods involving corrupted latent embeddings with noise injection for image reconstruction tasks.

---

5. Text-to-Video Generation Based on Diffusion Model

URL: View paper

Brief Assessment

Text-to-Video Diffusion[59] focuses on video generation using diffusion models with improved tokenizers, not on the specific training methodology of aligning tokenizer embeddings with denoising objectives through interpolative noise injection and random masking as proposed in the original paper.

---

6. Beyond Prompts: Preserving Semantics in Diffusion-based Communication

URL: View paper

Brief Assessment

Preserving Semantics Diffusion[62] focuses on semantic communication using textual inversion for image transmission over noisy channels, not on tokenizer training for generative models. The technical domains and objectives are fundamentally different.

---

7. Robust latent matters: Boosting image generation with sampling error synthesis

URL: View paper

Prior Art Analysis

Robust Latent Matters[56] demonstrates prior work on training tokenizers to reconstruct images from corrupted latent embeddings using noise injection. The candidate paper proposes 'latent perturbation' that corrupts latent embeddings by randomly replacing tokens with their nearest neighbors from the codebook, and trains the decoder to reconstruct original images from these perturbed latents. This approach directly addresses the same core problem: making tokenizers robust to corrupted/noisy latent representations during reconstruction. Both papers identify the training-inference discrepancy where decoders are trained on clean latents but must handle noisy latents during generation, and both propose injecting noise/perturbations into latent embeddings during tokenizer training to improve robustness.

Evidence

Evidence 1 - **Rationale**: Both papers use similar composite loss functions for tokenizer training, including reconstruction loss, perceptual loss, and adversarial loss, indicating similar training methodologies. - **Original**: our tokenizer is trained to reconstruct the original images from corrupted latent embeddings. the training objective follows established practice (rombach et al., 2022; esser et al., 2021; yu et al., 2025a), combining pixel-wise mean-squared-error (mse), latent-space kl-regularization (pu et al., 20... - **Candidate**: the robusttok is trained with composite losses including reconstruction loss lrecon, vector quantization loss lv q [15], adverserial losslad [24], perceptual losslp [27], and semantic loss lclip [35]

Evidence 2 - **Rationale**: Both papers identify the same fundamental problem: the discrepancy between tokenizer training (clean latents) and inference (potentially noisy latents), and propose addressing it by training tokenizers on corrupted latents. - **Original**: we observe that modern generative models share a conceptually similar training objective-reconstructing clean signals from corrupted inputs, such as signals degraded by gaussian noise or masking-a process we termdenoising. motivated by this insight, we propose aligning tokenizer embeddings directly ... - **Candidate**: knowing that ar models are subjected to sampling error accumulation due to the discrepancy between training and inference, we show in the following that such sampling error of ar models can be captured within the tokenizer alone with a novel reconstruction metric, and can be mitigated by involving p...

Evidence 3 - **Rationale**: Both papers describe corrupting latent embeddings during tokenizer training. The original uses 'interpolative noise' while the candidate uses 'latent perturbation', but the core mechanism of training decoders on corrupted latents is identical. - **Original**: during tokenizer training, we corrupt latent embeddings viainterpolative noise, generated by interpolating original embeddings with gaussian noise. the tokenizer decoder is then trained to reconstruct clean images from these heavily noised latent embeddings. - **Candidate**: during tokenizer training, we apply latent perturbation to enhance its robustness. we apply perturbation after semantic regularization [35] to preserve clear semantics in the discrete tokens to maximize the reconstruction capability. within a batch of image, we randomly choose β of them to add pertu...

---

8. GloTok: Global Perspective Tokenizer for Image Reconstruction and Generation

URL: View paper

Brief Assessment

GloTok[60] focuses on learning uniform semantic distributions through global histogram relations from pre-trained models, not on training tokenizers to reconstruct from corrupted latent embeddings via noise injection as in the original paper.

---

9. Discovering Latent Information from Noisy Sources

URL: View paper

Brief Assessment

Discovering Latent Information[61] focuses on extracting information from noisy social media data (text/images) for tasks like named entity recognition and domain adaptation, not on training tokenizers for generative models through latent embedding corruption and reconstruction.

---

10. Reduce information loss in transformers for pluralistic image inpainting

URL: View paper

Brief Assessment

Pluralistic Image Inpainting[58] focuses on image inpainting using patch-based tokenization without quantization, not on training tokenizers to reconstruct from corrupted latent embeddings using noise injection for generative modeling.

Contribution 2: Denoising as a unifying design principle for tokenizers

Description: The authors establish that modern generative models share a common training objective of reconstructing clean signals from corrupted inputs (denoising), and propose that tokenizers should be designed to align with this principle. This conceptual framework motivates tokenizer embeddings that remain reconstructable under significant corruption.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Generative Recommendation with Continuous-Token Diffusion

URL: View paper

Brief Assessment

Continuous-Token Diffusion[68] focuses on continuous tokenization for recommender systems using diffusion models, not on establishing denoising as a general design principle for tokenizers across generative models. The candidate addresses a different domain (recommendation) with different technical objectives.

---

2. Graph Diffusion Transformers are In-Context Molecular Designers

URL: View paper

Brief Assessment

Graph Diffusion Transformers[69] focuses on molecular design using demonstration-conditioned diffusion models and develops a molecular tokenizer with node pair encoding for motif-level representation. This is a domain-specific application in chemistry, not a general framework for tokenizer design principles across visual generative models.

---

3. TSG-DDT: Time-Series Generative Denoising Diffusion Transformers

URL: View paper

Brief Assessment

TSG-DDT[74] applies denoising diffusion to time-series MEA data generation for brain stimulation, not to tokenizer design for visual generative models. The denoising here refers to extracting noise from signals, not aligning tokenizer embeddings with downstream generative objectives.

---

4. BAMM: Bidirectional Autoregressive Motion Model

URL: View paper

Brief Assessment

BAMM[70] focuses on text-to-motion generation using a motion tokenizer and masked self-attention transformer, not on establishing denoising as a design principle for tokenizers in visual generative models. The domains and objectives are fundamentally different.

---

5. Rdpm: Solve diffusion probabilistic models via recurrent token prediction

URL: View paper

Brief Assessment

Rdpm[71] focuses on discrete diffusion through recurrent token prediction for image generation, not on designing tokenizers with denoising objectives. The candidate's tokenization is diffusion-based quantization, whereas the original proposes training tokenizers explicitly for denoising alignment with downstream generative models.

---

6. Masked autoencoders are effective tokenizers for diffusion models

URL: View paper

Brief Assessment

Autoencoders Diffusion Tokenizers[67] focuses on masked autoencoders for learning discriminative latent spaces in diffusion models, not on establishing denoising as a unifying design principle across generative model architectures. The candidate emphasizes mask modeling and latent space structure rather than proposing denoising as a fundamental tokenizer design principle.

---

7. Point-RTD: Replaced Token Denoising for Pretraining Transformer Models on Point Clouds

URL: View paper

Brief Assessment

Point-RTD[73] focuses on point cloud tokenization using replaced token denoising for 3D data, not on establishing denoising as a general design principle for tokenizers across generative models. The candidate addresses a specific domain (point clouds) rather than the broader conceptual framework proposed in the original paper.

---

8. Generalized Denoising Diffusion Codebook Models (gDDCM): Tokenizing images using a pre-trained diffusion model

URL: View paper

Brief Assessment

gDDCM[72] focuses on using pre-trained diffusion models for image tokenization through a denoising-backtracing sampling strategy, rather than proposing denoising as a general design principle for tokenizer development across diverse generative models.

---

9. Denoising token prediction in masked autoregressive models

URL: View paper

Brief Assessment

Denoising Token Prediction[65] focuses on applying denoising mechanisms within masked autoregressive models for token prediction during generation, not on designing tokenizers themselves. The candidate's denoising head refines predicted tokens during the generative process, whereas the original paper proposes denoising as a principle for training tokenizer embeddings to be robust under corruption.

---

10. Comparison of Autoencoders for tokenization of ASL datasets

URL: View paper

Brief Assessment

ASL Tokenization[66] focuses on comparing autoencoder architectures (feedforward, convolutional, diffusion) for ASL image tokenization, not on establishing denoising as a general design principle for tokenizers in generative models. The candidate's diffusion autoencoder uses denoising for noise robustness in a specific application domain, rather than proposing denoising alignment with downstream generative model objectives as a fundamental tokenizer design principle.

Contribution 3: Comprehensive empirical validation across diverse generative models

Description: The authors demonstrate that their tokenizer generalizes across six representative generative models (both autoregressive and non-autoregressive), multiple tokenizer architectures (2D continuous, 1D continuous, and vector-quantized), and different generation tasks, showing consistent improvements without requiring semantics distillation from external pretrained models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Vec-tok speech: speech vectorization and tokenization for neural speech generation

URL: View paper

Brief Assessment

Vec-tok Speech[54] focuses on speech generation tasks (voice conversion, text-to-speech, speech translation) using speech vectors and semantic tokens. The candidate does not address image generation or the specific tokenizer evaluation methodology across autoregressive and non-autoregressive visual generative models that the original paper claims as novel.

---

2. Genie: Generative interactive environments

URL: View paper

Brief Assessment

Genie[51] focuses on training generative interactive environments from unlabelled internet videos without action labels, using a latent action model. The original paper evaluates tokenizers across six generative models (both autoregressive and non-autoregressive). These are fundamentally different research directions with distinct objectives and evaluation paradigms.

---

3. Holistic Tokenizer for Autoregressive Image Generation

URL: View paper

Brief Assessment

Holistic Tokenizer[39] focuses on a holistic-to-local tokenization scheme for autoregressive image generation, testing primarily on AR models (mar, randomar, rasterar) and some non-AR models (dit, sit, lightningdit). The original paper's contribution emphasizes validation across six representative models (both AR and non-AR) with multiple tokenizer architectures (2D continuous, 1D continuous, vector-quantized) without semantics distillation. While both papers evaluate multiple generative models, Holistic Tokenizer[39] does not demonstrate the same breadth of tokenizer architecture types or explicitly avoid semantics distillation as a core design principle.

---

4. Magvit: Masked generative video transformer

URL: View paper

Brief Assessment

Magvit[53] focuses on masked video generation transformers with a 3D tokenizer for video synthesis tasks. The candidate does not address tokenizer generalization across multiple autoregressive and non-autoregressive image generation models, which is the core claim of the original contribution.

---

5. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation

URL: View paper

Brief Assessment

Open-magvit2[12] focuses on visual tokenization for auto-regressive image generation, not on validating tokenizers across multiple generative model types (both AR and non-AR) as claimed in the original paper. The candidate does not demonstrate prior work that validates tokenizers across six representative generative models spanning different architectures and generation tasks.

---

6. Frequency autoregressive image generation with continuous tokens

URL: View paper

Brief Assessment

Frequency Autoregressive Generation[55] focuses on frequency-progressive autoregressive generation with continuous tokens, not on validating tokenizers across multiple generative model architectures. The candidate's experiments primarily evaluate their FAR paradigm rather than demonstrating tokenizer generalizability across diverse AR and non-AR frameworks.

---

7. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation

URL: View paper

Brief Assessment

Llama Image Generation[52] focuses on autoregressive image generation models (LlamaGen) and does not evaluate tokenizers across both autoregressive and non-autoregressive frameworks as claimed in the original paper. The candidate validates their tokenizer only within the autoregressive paradigm.

---

8. Tokenflow: Unified image tokenizer for multimodal understanding and generation

URL: View paper

Brief Assessment

Tokenflow[4] focuses on unified image tokenization for multimodal understanding and generation using dual-codebook architecture. It does not address the specific claim about validating tokenizers across six representative generative models (both autoregressive and non-autoregressive) without semantics distillation, which is the core novelty of the original paper's contribution.

---

9. UniTok: A Unified Tokenizer for Visual Generation and Understanding

URL: View paper

Brief Assessment

UniTok[2] focuses on unifying tokenizers for both generation and understanding tasks through multi-codebook quantization, rather than validating a single tokenizer across multiple generative model architectures. The paper does not demonstrate testing across six different generative models (both AR and non-AR) as the original paper does.

---

10. Xq-gan: An open-source image tokenization framework for autoregressive generation

URL: View paper

Brief Assessment

Xq-gan[33] focuses on image tokenization techniques (quantization methods like VQ, RQ, PQ, LFQ, BSQ) rather than tokenizer training objectives. The candidate evaluates tokenizers primarily on reconstruction quality and generation with VAR/AR models, not on the denoising-alignment principle central to the original paper's contribution.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Latent Denoising Makes Good Visual Tokenizers View paper
- [1] Learning to tokenize for generative retrieval View paper
- [2] UniTok: A Unified Tokenizer for Visual Generation and Understanding View paper
- [3] Muse: Text-To-Image Generation via Masked Generative Transformers View paper
- [4] Tokenflow: Unified image tokenizer for multimodal understanding and generation View paper
- [5] Image tokenizer needs post-training View paper
- [6] Visual Prompt Tuning for Generative Transfer Learning View paper
- [7] Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation View paper
- [8] VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation View paper
- [9] Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens View paper
- [10] Scalable Visual Tokenizers for Generative Modeling View paper
- [11] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders View paper
- [12] Open-magvit2: An open-source project toward democratizing auto-regressive visual generation View paper
- [13] Atoken: A unified tokenizer for vision View paper
- [14] Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors View paper
- [15] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation View paper
- [16] Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation View paper
- [17] X-Omni: Reinforcement Learning Makes Discrete Autoregressive Image Generative Models Great Again View paper
- [18] V2Flow: Unifying Visual Tokenization and Large Language Model Vocabularies for Autoregressive Image Generation View paper
- [19] Factorized visual tokenization and generation View paper
- [20] ILLUME+: Illuminating Unified MLLM with Dual Visual Tokenization and Diffusion Refinement View paper
- [21] OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving View paper
- [22] QLIP: Text-Aligned Visual Tokenization Unifies Auto-Regressive Multimodal Understanding and Generation View paper
- [23] Vision foundation models as effective visual tokenizers for autoregressive image generation View paper
- [24] Aligning Visual Foundation Encoders to Tokenizers for Diffusion Models View paper
- [25] WeTok: Powerful Discrete Tokenization for High-Fidelity Visual Reconstruction View paper
- [26] CogView: Mastering Text-to-Image Generation via Transformers View paper
- [27] Layton: Latent Consistency Tokenizer for 1024-pixel Image Reconstruction and Generation by 256 Tokens View paper
- [28] A Simple Contrastive Framework Of Item Tokenization For Generative Recommendation View paper
- [29] Stabilize the latent space for image autoregressive modeling: A unified perspective View paper
- [30] A Survey on Vision-Language-Action Models: An Action Tokenization Perspective View paper
- [31] MaskGIT: Masked Generative Image Transformer View paper
- [32] ImageFolder: Autoregressive Image Generation with Folded Tokens View paper
- [33] Xq-gan: An open-source image tokenization framework for autoregressive generation View paper
- [34] Language of Stains: Tokenization Enhances Multiplex Immunofluorescence and Histology Image Synthesis View paper
- [35] Ming-UniVision: Joint Image Understanding and Generation with a Unified Continuous Tokenizer View paper
- [36] LARP: Tokenizing Videos with a Learned Autoregressive Generative Prior View paper
- [37] VTBench: Evaluating Visual Tokenizers for Autoregressive Image Generation View paper
- [38] OmniTokenizer: A Joint Image-Video Tokenizer for Visual Generation View paper
- [39] Holistic Tokenizer for Autoregressive Image Generation View paper
- [40] Flow to the Mode: Mode-Seeking Diffusion Autoencoders for State-of-the-Art Image Tokenization View paper
- [41] ResiComp: Loss-Resilient Image Compression via Dual-Functional Masked Visual Token Modeling View paper
- [42] Learnings from scaling visual tokenizers for reconstruction and generation View paper
- [43] MovieDreamer: Hierarchical Generation for Coherent Long Visual Sequence View paper
- [44] Stylerdalle: Language-guided style transfer using a vector-quantized tokenizer of a large-scale generative model View paper
- [45] Empowering Backbone Models for Visual Text Generation with Input Granularity Control and Glyph-Aware Training View paper
- [46] Spectral image tokenizer View paper
- [47] MAGE: MAsked Generative Encoder to Unify Representation Learning and Image Synthesis View paper
- [48] BiGR: Harnessing Binary Latent Codes for Image Generation and Improved Visual Representation Capabilities View paper
- [49] Resurrect mask autoregressive modeling for efficient and scalable image generation View paper
- [50] Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization View paper
- [51] Genie: Generative interactive environments View paper
- [52] Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation View paper
- [53] Magvit: Masked generative video transformer View paper
- [54] Vec-tok speech: speech vectorization and tokenization for neural speech generation View paper
- [55] Frequency autoregressive image generation with continuous tokens View paper
- [56] Robust latent matters: Boosting image generation with sampling error synthesis View paper
- [57] X-lrm: X-ray large reconstruction model for extremely sparse-view computed tomography recovery in one second View paper
- [58] Reduce information loss in transformers for pluralistic image inpainting View paper
- [59] Text-to-Video Generation Based on Diffusion Model View paper

- [60] GloTok: Global Perspective Tokenizer for Image Reconstruction and Generation View paper
- [61] Discovering Latent Information from Noisy Sources View paper
- [62] Beyond Prompts: Preserving Semantics in Diffusion-based Communication View paper
- [63] Brain-Semantoks: Learning Semantic Tokens of Brain Dynamics with a Self-Distilled Foundation Model View paper
- [64] LacTok: Latent Consistency Tokenizer for High-resolution Image Reconstruction and Generation by 256 Tokens View paper
- [65] Denoising token prediction in masked autoregressive models View paper
- [66] Comparison of Autoencoders for tokenization of ASL datasets View paper
- [67] Masked autoencoders are effective tokenizers for diffusion models View paper
- [68] Generative Recommendation with Continuous-Token Diffusion View paper
- [69] Graph Diffusion Transformers are In-Context Molecular Designers View paper
- [70] BAMM: Bidirectional Autoregressive Motion Model View paper
- [71] Rdpm: Solve diffusion probabilistic models via recurrent token prediction View paper
- [72] Generalized Denoising Diffusion Codebook Models (gDDCM): Tokenizing images using a pre-trained diffusion model View paper
- [73] Point-RTD: Replaced Token Denoising for Pretraining Transformer Models on Point Clouds View paper
- [74] TSG-DDT: Time-Series Generative Denoising Diffusion Transformers View paper