Novelty Assessment Report

Paper: AToken: A Unified Tokenizer for Vision

PDF URL: https://openreview.net/pdf?id=a4fSF5pGJq

Authors: Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, Yinfei Yang

Venue: ICLR 2026 Conference Withdrawn Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

We present AToken, the first unified visual tokenizer that achieves both high-fidelity reconstruction and semantic understanding across images, videos, and 3D assets. Unlike existing tokenizers that specialize in either reconstruction or understanding for single modalities, AToken encodes these diverse visual inputs in a shared 4D latent space, optimizing without separate model designs. Specifically, we introduce a pure transformer architecture with 4D rotary position embeddings to process visual inputs of arbitrary resolutions and temporal durations. To ensure stable training, we introduce an adversarial-free training objective that combines perceptual and Gram matrix losses, achieving state-of-the-art reconstruction quality. By employing a progressive training curriculum, AToken gradually expands from single images, videos, and 3D, and supports both continuous and discrete latent tokens. AToken achieves 0.21 rFID with 82.2% ImageNet accuracy for images, 3.01 rFVD with 32.6% MSRVTT retrieval for videos, and 28.19 PSNR with 90.9% classification accuracy for 3D. In downstream applications, AToken enables both visual generation tasks (e.g., image generation with continuous and discrete tokens, text-to-video generation, image-to-3D synthesis) and understanding tasks (e.g., multimodal LLMs), achieving 1.44/2.23 gFID on ImageNet for continuous/discrete tokens, and 48.7% on MMMU and 64.5% on VideoMME. These results shed light on the next-generation multimodal AI systems built upon the unified visual tokenization.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **unified visual tokenization across images videos and 3D**

A total of **50 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

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

- **Unified Multimodal Tokenization Architectures**
- **Video-Specific Tokenization Methods**
- **3D Scene Tokenization and Understanding**
- **Token Reduction and Efficiency for Multimodal LLMs**
- **Multimodal Integration and Application-Driven Tokenization**
- **Specialized Tokenization for Downstream Tasks**
- **Foundational Concepts and Broad Surveys**

Complete Taxonomy Tree

- unified visual tokenization across images videos and 3D Survey Taxonomy
- Unified Multimodal Tokenization Architectures
  - Shared Latent Space Tokenizers ★ (3 papers)
  - [0] AToken: A Unified Tokenizer for Vision (Lu et al., 2026) View paper
  - [13] Show-o2: Improved Native Unified Multimodal Models (Xie, 2025) View paper
  - [15] Omnitokenizer: A joint image-video tokenizer for visual generation (Yi Jiang, 2024) View paper
  - Frozen Encoder Multimodal Frameworks (1 papers)
  - [4] Meta-transformer: A unified framework for multimodal learning (Zhang, 2023) View paper
  - Heterogeneous Signal Tokenization (1 papers)
  - [2] Harmonizer: A Universal Signal Tokenization Framework for Multimodal Large Language Models (Amin Amiri, 2025) View paper
- Video-Specific Tokenization Methods
  - Temporal Compression and Efficiency
  - Adaptive Token Allocation (3 papers)
    - [5] Efficient Long Video Tokenization via Coordinate-based Patch Reconstruction (Huiwon Jang, 2025) View paper
    - [23] Learning Adaptive and Temporally Causal Video Tokenization in a 1D Latent Space (Li Yan, 2025) View paper
    - [41] ElasticTok: Adaptive Tokenization for Image and Video (Yan, 2024) View paper
  - Coordinate-Based and Trajectory-Based Tokenization (1 papers)
    - [22] One Trajectory, One Token: Grounded Video Tokenization via Panoptic Sub-object Trajectory (Zheng Chenhao, 2025) View paper
  - Hierarchical and Progressive Tokenization (2 papers)
    - [39] HiTVideo: Hierarchical Tokenizers for Enhancing Text-to-Video Generation with Autoregressive Large Language Models (Zhou, 2025) View paper
    - [46] Progressive Growing of Video Tokenizers for Temporally Compact Latent Spaces (Mahapatra, 2025) View paper
  - Reconstruction-Oriented Video Tokenizers (3 papers)
  - [8] Rethinking video tokenization: A conditioned diffusion-based approach (Yang Nianzu, 2025) View paper

  - [11] Image and video tokenization with binary spherical quantization (Zhao Yue, 2024) View paper
  - [18] Vidtok: A versatile and open-source video tokenizer (Tang, 2024) View paper
  - Semantic and Spatiotemporal Decoupling (3 papers)
  - [10] Sweettok: Semantic-aware spatial-temporal tokenizer for compact video discretization (Tan, 2025) View paper
  - [14] Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization (Jin Yang, 2024) View paper
  - [30] Versatile Video Tokenization with Generative 2D Gaussian Splatting (Chen Zhenghao, 2025) View paper
  - Specialized Video Tokenization Applications (2 papers)
  - [24] Resi-VidTok: An Efficient and Decomposed Progressive Tokenization Framework for Ultra-Low-Rate and Lightweight Video Transmission (Liu Zhen-yu, 2025) View paper
  - [28] Video Segmentation and Tokenization for Model-Based Video Scene Classification (Qing Wang, 2025) View paper
- 3D Scene Tokenization and Understanding
  - 3D Scene Forecasting and World Models (1 papers)
  - [3] I2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting (Liao Zhimin, 2025) View paper
  - 3D-Aware Multimodal LMMs (2 papers)
  - [16] LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness (Zhu, 2024) View paper
  - [17] VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction (Fan Zhiwen, 2025) View paper
  - Cross-Modal 2D-3D Tokenization Alignment (2 papers)
  - [27] Sam-guided masked token prediction for 3d scene understanding (Zhimin Chen, 2024) View paper
  - [37] TDFormer: Top-Down Token Generation for 3D Medical Image Segmentation. (Hao Du, 2025) View paper
- Token Reduction and Efficiency for Multimodal LLMs
  - Long Video Understanding with Token Compression (3 papers)
  - [31] STORM: Token-Efficient Long Video Understanding for Multimodal LLMs (Jiang Jindong, 2025) View paper
  - [34] Video-XL-Pro: Reconstructive Token Compression for Extremely Long Video Understanding (Liu Xiang-rui, 2025) View paper
  - Spatial Token Reduction for Images and Videos (3 papers)
  - [25] Tokenlearner: Adaptive space-time tokenization for videos (Michael S. Ryoo, 2021) View paper
  - [32] AdaToken-3D: Dynamic Spatial Gating for Efficient 3D Large Multimodal-Models Reasoning (Zhang Kai, 2025) View paper
  - [33] Video Token Sparsification for Efficient Multimodal LLMs in Driving Visual Question Answering (Yunsheng MA, 2025) View paper
  - Segment-Aware and Task-Specific Token Optimization (1 papers)
  - [44] SAGE: Segment-Aware Gloss-Free Encoding for Token-Efficient Sign Language Translation (Sincan, 2025) View paper
- Multimodal Integration and Application-Driven Tokenization
  - Vision-Language Models for Dense Grounding and Segmentation (3 papers)
  - [7] Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos (Yuan, 2025) View paper
  - [40] Chain-of-Visual-Thought: Teaching VLMs to See and Think Better with Continuous Visual Tokens (Yiming Qin, 2025) View paper
  - [48] Grounding Everything in Tokens for Multimodal Large Language Models (Xiangxuan Ren, 2025) View paper
  - Multi-Image and Multi-Frame Multimodal LMMs (2 papers)
  - [1] LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models (Li Feng, 2024) View paper
  - [42] TinyLLaVA-Video: Towards Smaller LMMs for Video Understanding with Group Resampler (Zhang Xingjian, 2025) View paper
  - Medical and Domain-Specific Multimodal Tokenization (2 papers)
  - [20] OmniV-Med: Scaling Medical Vision-Language Model for Universal Visual Understanding (Wang Yuan, 2025) View paper
  - [45] Better Tokens for Better 3D: Advancing Vision-Language Modeling in 3D Medical Imaging (Hamamci, 2025) View paper
  - Vision-Language-Action and Embodied AI Tokenization (3 papers)
  - [6] Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts (Chuan Guo, 2022) View paper
  - [19] Being-h0: vision-language-action pretraining from large-scale human videos (Luo Hao, 2025) View paper
  - [21] UniPose: A Unified Multimodal Framework for Human Pose Comprehension, Generation and Editing (Yiheng Li, 2024) View paper
  - Cross-Modal Translation and Generation (2 papers)
  - [26] Vx2text: End-to-end learning of video-based text generation from multimodal inputs (Xudong Lin, 2021) View paper
  - [35] Unified Cross-modal Translation of Score Images, Symbolic Music, and Performance Audio (Jung Jong-Min, 2025) View paper
- Specialized Tokenization for Downstream Tasks
  - Video Question Answering and Reasoning (2 papers)
  - [29] Video question answering with iterative video-text co-tokenization (AJ Piergiovanni, 2022) View paper
  - [38] Leveraging Vision-Language Large Models for Interpretable Video Action Recognition with Semantic Tokenization (Peng Jingwei, 2025) View paper
  - Action Recognition and Behavior Understanding (2 papers)
  - [9] Shopformer: Transformer-Based Framework for Detecting Shoplifting via Human Pose (Narges Rashvand, 2025) View paper
  - [36] Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition (Z Gao, 2015) View paper
  - Object Detection and Tracking with Token Sequences (2 papers)
  - [47] Object Discovery in Images, Videos, and 3D Scenes (Yangtao, 2024) View paper
  - [50] Improving Token-based Object Detection with Video (Abhineet Singh, 2025) View paper
  - Non-Visual Domain Tokenization (1 papers)
  - [12] Prot2token: A multi-task framework for protein language processing using autoregressive language modeling (Mahdi Pourmirzaei, 2024) View paper
- Foundational Concepts and Broad Surveys (1 papers)
  - [49] Visual Content Synthesis at Scale (Ge, 2025) View paper

Narrative

Core task: unified visual tokenization across images, videos, and 3D. The field has evolved to address the challenge of representing diverse visual modalities—static images, temporal video sequences, and spatial 3D structures—within a common tokenization framework that can interface with large language models and generative systems. The taxonomy reveals several complementary directions: Unified Multimodal Tokenization Architectures pursue shared latent spaces that handle multiple modalities through common encoders or cross-modal alignment (e.g., Omnitokenizer[15], Show-o2[13]); Video-Specific Tokenization Methods focus on temporal compression and causal modeling tailored to video data (Vidtok[18], Causal Video Tokenization[23]); 3D Scene Tokenization and Understanding develops representations for point clouds, meshes, and volumetric data (LLaVA-3D[16], Gaussian Splatting Tokenization[30]); Token Reduction and Efficiency branches explore adaptive pruning and sparsification to manage computational costs in multimodal LLMs (ElasticTok[41], AdaToken-3D[32]); Multimodal Integration and Application-Driven Tokenization emphasizes end-to-end systems for tasks like visual question answering and content synthesis (LLaVA-NeXT-Interleave[1], I2-world[3]); and Specialized Tokenization for Downstream Tasks targets domain-specific needs such as action recognition or medical imaging (Semantic Action Tokenization[38], OmniV-Med[20]).

Within the Unified Multimodal Tokenization Architectures branch, a central theme is whether to learn a single shared encoder or maintain modality-specific pathways that converge in a common latent space. AToken[0] exemplifies the shared latent space approach, aiming to unify image, video, and 3D representations through a cohesive tokenization strategy that balances expressiveness across modalities. This contrasts with works like Meta-transformer[4], which employ modality-specific preprocessing before a unified transformer backbone, and Harmonizer[2], which focuses on aligning heterogeneous token distributions post-encoding. Compared to neighbors such as Show-o2[13], which integrates discrete codebook learning for joint image-text generation, and Omnitokenizer[15], which emphasizes cross-modal retrieval and alignment, AToken[0] prioritizes a more tightly coupled latent space that directly supports reasoning and generation across all three visual domains. The trade-offs revolve around reconstruction fidelity, computational overhead, and the ease of extending tokenization to new modalities or downstream tasks, with ongoing questions about optimal codebook sizes, temporal modeling granularity, and the role of 3D geometric priors in unified frameworks.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Show-o2: Improved Native Unified Multimodal Models

Authors: Xie, Jinheng, Yang Zhenheng, Jinheng Xie, Shou, et al. (8 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

This paper presents improved native unified multimodal models, \emph{i.e.,} Show-o2, that leverage autoregressive modeling and flow matching. Built upon a 3D causal variational autoencoder space, unified visual representations are constructed through a dual-path of spatial (-temporal) fusion, enabling scalability across image and video modalities while ensuring effective multimodal understanding and generation. Based on a language model, autoregressive modeling and flow matching are natively app...

Relationship Analysis

Both papers belong to the Shared Latent Space Tokenizers category, encoding diverse visual inputs (images, videos, 3D) into unified representations for cross-modal processing. They overlap in their goal of unified multimodal tokenization and support for both understanding and generation tasks across visual modalities. However, AToken uses a sparse 4D transformer architecture with adversarial-free training and progressive curriculum learning, while Show-o2 employs a 3D causal VAE with dual-path spatial-temporal fusion and combines autoregressive modeling with flow matching for generation.

2. Omnitokenizer: A joint image-video tokenizer for visual generation

Authors: Yi Jiang, Yu-Gang Jiang, Binyue Peng, Junke Wang, Zuxuan Wu, et al. (6 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ video inputs, this paper presents OmniTokenizer, a transformer-based tokenizer for joint image and video â¦ Although MAGVITv2 [67] have explored causal 3D convolution to process both â¦

Relationship Analysis

Both papers belong to the Shared Latent Space Tokenizers category, encoding diverse visual inputs into unified representations for cross-modal processing. They overlap in addressing unified tokenization across images and videos using transformer architectures with spatial-temporal modeling. However, AToken extends to 3D assets using 4D representations with rotary position embeddings and achieves joint reconstruction-understanding capabilities, while OmniTokenizer focuses solely on image-video tokenization with a spatial-temporal decoupled architecture and progressive training from images to videos without 3D support or semantic understanding objectives.

Contributions Analysis

Overall novelty summary. AToken proposes a unified visual tokenizer that encodes images, videos, and 3D assets into a shared 4D latent space, targeting both high-fidelity reconstruction and semantic understanding. The paper resides in the 'Shared Latent Space Tokenizers' leaf, which contains only three papers including AToken itself. This leaf sits within the broader 'Unified Multimodal Tokenization Architectures' branch, indicating a relatively sparse but emerging research direction. The small sibling count suggests that truly unified tokenizers handling all three modalities (images, videos, 3D) within a single architecture remain uncommon, positioning AToken in a less crowded area of the field.

The taxonomy reveals that most related work either specializes in single modalities or adopts modality-specific preprocessing before unification. The 'Video-Specific Tokenization Methods' branch contains numerous papers focused solely on temporal compression and reconstruction, while '3D Scene Tokenization and Understanding' addresses point clouds and volumetric data separately. Neighboring leaves like 'Frozen Encoder Multimodal Frameworks' and 'Heterogeneous Signal Tokenization' pursue cross-modal alignment through different architectural strategies—frozen pretrained encoders or discrete token conversion for LLMs—rather than AToken's approach of learning a shared continuous latent space from scratch across all three visual domains.

Among 29 candidates examined, the contribution-level analysis reveals mixed novelty signals. The core unified tokenizer concept (Contribution A) examined 9 candidates with no clear refutations, suggesting limited direct prior work on this specific three-modality unification. The 4D rotary position embeddings (Contribution B) also found no refutations across 10 candidates, indicating architectural novelty. However, the adversarial-free training objective with Gram matrix loss (Contribution C) encountered 1 refutable candidate among 10 examined, pointing to some overlap with existing reconstruction training strategies. The limited search scope means these findings reflect top-30 semantic matches rather than exhaustive coverage.

Given the sparse taxonomy leaf and limited refutations across most contributions, AToken appears to occupy a relatively novel position within the examined literature. The main uncertainty concerns whether the adversarial-free training approach represents a significant departure from prior reconstruction methods, and whether the 30-candidate search captured all relevant unified tokenization work. The analysis suggests incremental innovation in training objectives but potentially stronger novelty in the architectural unification of three visual modalities within a single learned latent space.

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

Contribution 1: AToken: unified visual tokenizer for images, videos, and 3D

Description: The authors introduce AToken, a single tokenizer that unifies reconstruction and understanding tasks across three visual modalities (images, videos, 3D assets) by encoding them into a shared 4D latent space, supporting both continuous and discrete token representations.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Image and video tokenization with binary spherical quantization

URL: View paper

Brief Assessment

Binary Spherical Quantization[11] focuses on image and video tokenization using binary spherical quantization with transformer architectures, but does not address 3D assets or unified understanding tasks across modalities.

2. MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation

URL: View paper

Brief Assessment

MTVCrafter[55] focuses on 4D motion tokenization specifically for human image animation tasks, not on unified visual tokenization across images, videos, and 3D assets for both reconstruction and understanding.

3. Factorized visual tokenization and generation

URL: View paper

Brief Assessment

Factorized Visual Tokenization[54] focuses on factorized quantization for image tokenization and generation, not on unifying multiple visual modalities (images, videos, 3D) into a shared representation space as AToken does.

4. Omnitokenizer: A joint image-video tokenizer for visual generation

URL: View paper

Brief Assessment

Omnitokenizer[15] focuses exclusively on joint image-video tokenization without 3D capabilities. The paper explicitly states it handles 'both image and video inputs within a unified framework' but makes no mention of 3D assets, which is a core component of the original paper's contribution.

5. Advanced sign language video generation with compressed and quantized multi-condition tokenization

URL: View paper

Brief Assessment

Sign Language Tokenization[58] focuses on sign language video generation using multi-condition tokenization (poses, 3D hands) for a specific domain application, not on creating a unified visual tokenizer across general image, video, and 3D modalities for both reconstruction and understanding tasks.

6. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token

URL: View paper

Brief Assessment

LLaVA-Mini[56] focuses on efficient vision token compression for multimodal LLMs (reducing vision tokens to 1), not on unified visual tokenization across modalities for both reconstruction and understanding tasks.

7. Learnings from scaling visual tokenizers for reconstruction and generation

URL: View paper

Brief Assessment

Scaling Visual Tokenizers[51] focuses on scaling transformer-based auto-encoders (vitok) for image and video reconstruction/generation, exploring bottleneck sizes and encoder/decoder scaling. It does not address 3D assets or unified multimodal tokenization across images, videos, and 3D in a single framework with both reconstruction and understanding capabilities.

8. Language Model Beats Diffusion--Tokenizer is Key to Visual Generation

URL: View paper

Brief Assessment

Tokenizer Visual Generation[53] focuses on a video tokenizer (MAGVIT-v2) that handles images and videos with a shared vocabulary, but does not address 3D assets. The original paper's novelty lies in unifying all three modalities (images, videos, and 3D) in a single tokenizer framework with 4D latent space representation.

9. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies

URL: View paper

Brief Assessment

Dualtoken[57] focuses on unifying visual understanding and generation through dual visual vocabularies (semantic and perceptual codebooks) for images only, not extending to videos or 3D assets. The candidate does not address multi-modal tokenization across images, videos, and 3D in a shared 4D latent space.

Contribution 2: Pure transformer architecture with 4D rotary position embeddings

Description: The authors propose a transformer-based encoder-decoder architecture that extends 2D image processing to a unified 4D space using rotary position embeddings, enabling native handling of arbitrary resolutions and temporal lengths across all modalities.

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. Rotary Position Embedding for Vision Transformer

URL: View paper

Brief Assessment

Rotary Vision Transformer[60] focuses on applying 2D RoPE to standard vision transformers for image classification and segmentation tasks, not on unified 4D multimodal processing across images, videos, and 3D assets as in the original paper.

2. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution

URL: View paper

Brief Assessment

Qwen2-VL[64] uses multimodal rotary position embedding (m-rope) for 2D images and videos, not a unified 4D space representation (t,x,y,z) that includes 3D assets. The original paper's 4D rope spans temporal and full 3D spatial dimensions for images, videos, and 3D geometry.

3. HoPE: Hybrid of Position Embedding for Length Generalization in Vision-Language Models

URL: View paper

Brief Assessment

HoPE[65] focuses on extending 2D RoPE to 3D video inputs (t,x,y) for vision-language models, while the original paper proposes a unified 4D space (t,x,y,z) for images, videos, and 3D assets with a pure transformer tokenizer architecture. These are fundamentally different architectural goals and application domains.

4. Revisiting Multimodal Positional Encoding in Vision-Language Models

URL: View paper

Brief Assessment

Multimodal Positional Encoding[59] focuses on analyzing and improving rotary positional embeddings for vision-language models, not proposing a unified 4D transformer architecture for visual tokenization across images, videos, and 3D assets.

5. Circle-RoPE: Cone-like Decoupled Rotary Positional Embedding for Large Vision-Language Models

URL: View paper

Brief Assessment

Circle-RoPE[63] focuses on modifying RoPE for vision-language models to reduce cross-modal positional biases, not on proposing a unified 4D transformer architecture for multimodal visual processing across images, videos, and 3D assets.

6. VideoRoPE: What Makes for Good Video Rotary Position Embedding?

URL: View paper

Brief Assessment

VideoRoPE[62] focuses on extending RoPE to video data with temporal-spatial structure for video LLMs, while the original paper presents a unified visual tokenizer architecture for reconstruction and understanding across images, videos, and 3D assets. These are fundamentally different architectural goals and applications.

7. Liere: Generalizing rotary position encodings

URL: View paper

Brief Assessment

Liere[67] focuses on generalizing rotary position encodings to higher dimensions (2D, 3D) for classification tasks, not on unified visual tokenization across modalities for both reconstruction and understanding as in the original paper.

8. Vrope: Rotary position embedding for video large language models

URL: View paper

Brief Assessment

Vrope[61] focuses on extending RoPE for video-LLMs with spatiotemporal encoding, not on unified multimodal tokenization across images, videos, and 3D assets with reconstruction capabilities as in the original paper.

9. Medical image interpretation with large multimodal models

URL: View paper

Brief Assessment

Medical Multimodal Interpretation[68] only mentions rotary embeddings in the context of a language model (phi-1.5) with standard architecture. It does not describe a transformer architecture with 4D rotary position embeddings for unified multimodal visual processing across images, videos, and 3D data as proposed in the original paper.

10. EVA02-AT: Egocentric Video-Language Understanding with Spatial-Temporal Rotary Positional Embeddings and Symmetric Optimization

URL: View paper

Brief Assessment

EVA02-AT[66] focuses on egocentric video-language understanding with spatial-temporal rope for video encoders, not a unified tokenizer for multimodal reconstruction and understanding across images, videos, and 3D assets.

Contribution 3: Adversarial-free training objective with Gram matrix loss

Description: The authors develop a stable training approach that replaces adversarial training with a combination of perceptual and Gram matrix losses, directly optimizing second-order statistics to achieve high-fidelity reconstruction without GAN instabilities.

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. A training method for image compression networks to improve perceptual quality of reconstructions

URL: View paper

Brief Assessment

Perceptual Quality Training[74] uses adversarial training with GAN-based discriminators combined with perceptual and VGG losses. The original paper explicitly avoids adversarial training, developing a stable alternative using Gram matrix loss to optimize second-order statistics without discriminators.

2. Auxiliary loss reweighting for image inpainting

URL: View paper

Brief Assessment

Auxiliary Loss Reweighting[77] focuses on reweighting perceptual and style losses for image inpainting, not on replacing adversarial training. The candidate uses Gram matrix loss as one component among multiple auxiliary losses but does not propose it as an adversarial-free alternative for general reconstruction tasks.

3. From Images to Perception: Emergence of Perceptual Properties by Reconstructing Images

URL: View paper

Brief Assessment

Perceptual Properties Emergence[78] focuses on self-supervised learning of perceptual representations through image reconstruction tasks in biologically-inspired architectures, not on adversarial-free training methods using Gram matrix losses for visual tokenization.

4. SSDD: Single-Step Diffusion Decoder for Efficient Image Tokenization

URL: View paper

Brief Assessment

SSDD[71] focuses on diffusion-based decoders for image tokenization and mentions 'gan-free training' but does not describe using Gram matrix loss or the specific combination of perceptual and Gram matrix losses for optimizing second-order statistics as proposed in the original paper.

5. Flexible style image super-resolution using conditional objective

URL: View paper

Brief Assessment

Conditional Style Super-Resolution[76] uses perceptual losses at different feature levels for style-controllable SR, but does not specifically employ Gram matrix loss or address adversarial training instabilities in the context of visual tokenization across multiple modalities.

6. Local Texture Pattern Estimation for Image Detail Super-Resolution

URL: View paper

Prior Art Analysis

Local Texture Pattern[70] demonstrates that adversarial-free training using L1 loss and Gram loss for texture reconstruction was already established prior to the original paper. The candidate explicitly states that their method uses 'l1 loss and gram loss are simultaneously used to optimize the network' to 'effectively recover high-frequency texture without using gan structures.' This directly shows that combining perceptual losses with Gram matrix loss for adversarial-free training was not novel to the original paper, as Local Texture Pattern[70] had already implemented this approach for image super-resolution tasks.

Evidence

Evidence 1 - **Rationale**: Both papers describe adversarial-free training approaches that combine pixel-level losses (L1) with Gram matrix loss to achieve high-quality reconstruction without GAN instabilities. Local Texture Pattern[70] explicitly states this combination was used to optimize their network for texture recovery without GANs. - **Original**: to overcome training instabilities that affect transformer-based visual tokenizers, we develop an adversarial-free loss combining perceptual and gram matrix terms. this approach achieves state-ofthe-art reconstruction quality while maintaining stable, scalable training. - **Candidate**: finally, l1 loss and gram loss are simultaneously used to optimize the network. experimental results demonstrate that the proposed method can effectively recover high-frequency texture without using gan structures.

Evidence 2 - **Rationale**: Both papers emphasize using Gram matrix-based approaches to avoid adversarial training. Local Texture Pattern[70] demonstrates that adversarial-free texture reconstruction using Gram loss was already an established approach before the original paper's submission. - **Original**: this motivated adopting gram matrix loss (gatys et al., 2016), which directly optimizes feature covariance without adversarial training by computing the gram matrixg(f)= ff > for feature maps from different layers. - **Candidate**: inspired by traditional texture analysis research, this paper proposes a novel sr network based on local texture pattern estimation (ltpe), which can restore fine high-frequency texture details without gan.

7. MR image reconstruction using deep learning: evaluation of network structure and loss functions

URL: View paper

Brief Assessment

MR Deep Learning[75] focuses on MR image reconstruction using perceptual loss for medical imaging, not on adversarial-free training with Gram matrix loss for visual tokenization across multiple modalities.

8. Single image HDR reconstruction using a CNN with masked features and perceptual loss

URL: View paper

Brief Assessment

HDR Masked Features[72] focuses on single image HDR reconstruction using perceptual and Gram matrix losses, but does not address the specific context of unified visual tokenization across multiple modalities (images, videos, 3D) that the original paper targets. The application domains and technical challenges are fundamentally different.

9. MFMAM: Image inpainting via multi-scale feature module with attention module

URL: View paper

Brief Assessment

MFMAM[69] uses style loss and perceptual loss for image inpainting consistency, but does not describe an adversarial-free training approach or specifically employ Gram matrix loss as a replacement for GAN training. The candidate focuses on inpainting architecture rather than tokenizer training methodology.

10. Brain-driven facial image reconstruction via StyleGAN inversion with improved identity consistency

URL: View paper

Brief Assessment

Brain-driven StyleGAN[73] focuses on fMRI-to-face reconstruction using identity loss and custom loss functions for facial features, not on adversarial-free training with Gram matrix losses for general visual reconstruction.

Appendix: Text Similarity Detection

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

References

- [0] AToken: A Unified Tokenizer for Vision View paper
- [1] LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models View paper
- [2] Harmonizer: A Universal Signal Tokenization Framework for Multimodal Large Language Models View paper

- [3] I2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting View paper
- [4] Meta-transformer: A unified framework for multimodal learning View paper
- [5] Efficient Long Video Tokenization via Coordinate-based Patch Reconstruction View paper
- [6] Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts View paper
- [7] Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos View paper
- [8] Rethinking video tokenization: A conditioned diffusion-based approach View paper
- [9] Shopformer: Transformer-Based Framework for Detecting Shoplifting via Human Pose View paper
- [10] Sweettok: Semantic-aware spatial-temporal tokenizer for compact video discretization View paper
- [11] Image and video tokenization with binary spherical quantization View paper
- [12] Prot2token: A multi-task framework for protein language processing using autoregressive language modeling View paper
- [13] Show-o2: Improved Native Unified Multimodal Models View paper
- [14] Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization View paper
- [15] Omnitokenizer: A joint image-video tokenizer for visual generation View paper
- [16] LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness View paper
- [17] VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction View paper
- [18] Vidtok: A versatile and open-source video tokenizer View paper
- [19] Being-h0: vision-language-action pretraining from large-scale human videos View paper
- [20] OmniV-Med: Scaling Medical Vision-Language Model for Universal Visual Understanding View paper
- [21] UniPose: A Unified Multimodal Framework for Human Pose Comprehension, Generation and Editing View paper
- [22] One Trajectory, One Token: Grounded Video Tokenization via Panoptic Sub-object Trajectory View paper
- [23] Learning Adaptive and Temporally Causal Video Tokenization in a 1D Latent Space View paper
- [24] Resi-VidTok: An Efficient and Decomposed Progressive Tokenization Framework for Ultra-Low-Rate and Lightweight Video Transmission View paper
- [25] Tokenlearner: Adaptive space-time tokenization for videos View paper
- [26] Vx2text: End-to-end learning of video-based text generation from multimodal inputs View paper
- [27] Sam-guided masked token prediction for 3d scene understanding View paper
- [28] Video Segmentation and Tokenization for Model-Based Video Scene Classification View paper
- [29] Video question answering with iterative video-text co-tokenization View paper
- [30] Versatile Video Tokenization with Generative 2D Gaussian Splatting View paper
- [31] STORM: Token-Efficient Long Video Understanding for Multimodal LLMs View paper
- [32] AdaToken-3D: Dynamic Spatial Gating for Efficient 3D Large Multimodal-Models Reasoning View paper
- [33] Video Token Sparsification for Efficient Multimodal LLMs in Driving Visual Question Answering View paper
- [34] Video-XL-Pro: Reconstructive Token Compression for Extremely Long Video Understanding View paper
- [35] Unified Cross-modal Translation of Score Images, Symbolic Music, and Performance Audio View paper
- [36] Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition View paper
- [37] TDFormer: Top-Down Token Generation for 3D Medical Image Segmentation. View paper
- [38] Leveraging Vision-Language Large Models for Interpretable Video Action Recognition with Semantic Tokenization View paper
- [39] HiTVideo: Hierarchical Tokenizers for Enhancing Text-to-Video Generation with Autoregressive Large Language Models View paper
- [40] Chain-of-Visual-Thought: Teaching VLMs to See and Think Better with Continuous Visual Tokens View paper
- [41] ElasticTok: Adaptive Tokenization for Image and Video View paper
- [42] TinyLLaVA-Video: Towards Smaller LMMs for Video Understanding with Group Resampler View paper
- [43] Token-Efficient Long Video Understanding for Multimodal LLMs View paper
- [44] SAGE: Segment-Aware Gloss-Free Encoding for Token-Efficient Sign Language Translation View paper
- [45] Better Tokens for Better 3D: Advancing Vision-Language Modeling in 3D Medical Imaging View paper
- [46] Progressive Growing of Video Tokenizers for Temporally Compact Latent Spaces View paper
- [47] Object Discovery in Images, Videos, and 3D Scenes View paper
- [48] Grounding Everything in Tokens for Multimodal Large Language Models View paper
- [49] Visual Content Synthesis at Scale View paper
- [50] Improving Token-based Object Detection with Video View paper
- [51] Learnings from scaling visual tokenizers for reconstruction and generation View paper
- [52] UniFlow: A Unified Pixel Flow Tokenizer for Visual Understanding and Generation View paper
- [53] Language Model Beats Diffusion--Tokenizer is Key to Visual Generation View paper
- [54] Factorized visual tokenization and generation View paper
- [55] MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation View paper
- [56] LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token View paper
- [57] Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies View paper
- [58] Advanced sign language video generation with compressed and quantized multi-condition tokenization View paper
- [59] Revisiting Multimodal Positional Encoding in Vision-Language Models View paper
- [60] Rotary Position Embedding for Vision Transformer View paper
- [61] Vrope: Rotary position embedding for video large language models View paper
- [62] VideoRoPE: What Makes for Good Video Rotary Position Embedding? View paper
- [63] Circle-RoPE: Cone-like Decoupled Rotary Positional Embedding for Large Vision-Language Models View paper
- [64] Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution View paper
- [65] HoPE: Hybrid of Position Embedding for Length Generalization in Vision-Language Models View paper
- [66] EVA02-AT: Egocentric Video-Language Understanding with Spatial-Temporal Rotary Positional Embeddings and Symmetric Optimization View paper
- [67] Liere: Generalizing rotary position encodings View paper
- [68] Medical image interpretation with large multimodal models View paper
- [69] MFMAM: Image inpainting via multi-scale feature module with attention module View paper
- [70] Local Texture Pattern Estimation for Image Detail Super-Resolution View paper
- [71] SSDD: Single-Step Diffusion Decoder for Efficient Image Tokenization View paper

- [72] Single image HDR reconstruction using a CNN with masked features and perceptual loss View paper
- [73] Brain-driven facial image reconstruction via StyleGAN inversion with improved identity consistency View paper
- [74] A training method for image compression networks to improve perceptual quality of reconstructions View paper
- [75] MR image reconstruction using deep learning: evaluation of network structure and loss functions View paper
- [76] Flexible style image super-resolution using conditional objective View paper
- [77] Auxiliary loss reweighting for image inpainting View paper
- [78] From Images to Perception: Emergence of Perceptual Properties by Reconstructing Images View paper