Novelty Assessment Report

Paper: Adapting Self-Supervised Representations as a Latent Space for Efficient Generation
PDF URL: https://openreview.net/pdf?id=0b6a2SE23v
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

We introduce Representation Tokenizer (RepTok), a generative modeling framework that represents an image using a single continuous latent token obtained from self-supervised vision transformers. Building on a pre-trained SSL encoder, we fine-tune only the semantic token embedding and pair it with a generative decoder trained end-to-end using a standard flow matching objective. This adaptation enriches the token with low-level, reconstruction-relevant details, enabling faithful image reconstruction. To preserve the favorable geometry of the original SSL space, we add a cosine-similarity loss that regularizes the adapted token, ensuring it remains smooth and suitable for generation.Our single-token formulation resolves the spatial redundancies of the 2D latent space, simplifies architectures, and significantly reduces training costs. Despite its simplicity and efficiency, RepTok achieves competitive results on class-conditional ImageNet generation and extends naturally to text-to-image synthesis, reaching competitive zero-shot performance on MS-COCO under extremely limited training budgets. Our findings highlight the potential of fine-tuned SSL representations as compact and effective latent spaces for efficient generative modeling. We will release our model to facilitate further research.

Core Task Landscape

This paper addresses: **Efficient Image Generation from Single Continuous Latent Token**
A total of **50 papers** were analyzed and organized into a taxonomy with **33 categories**.

Taxonomy Overview

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

- **Continuous Token Representation and Tokenization**
- **Autoregressive Generation with Continuous Tokens**
- **Hybrid and Discrete-Continuous Approaches**
- **Unified Multimodal Generation and Understanding**
- **Diffusion-Based Generation in Continuous Latent Spaces**
- **Conditional and Controllable Generation**
- **Domain-Specific and Application-Oriented Generation**
- **Representation Learning and Latent Space Properties**
- **Cross-Modal and Multimodal Synthesis**
- **Auxiliary Techniques and Architectural Components**

Complete Taxonomy Tree

- Efficient Image Generation from Single Continuous Latent Token Survey Taxonomy
- Continuous Token Representation and Tokenization
  - Single-Token and Ultra-Compact Representations ★ (2 papers)
  - [0] Adapting Self-Supervised Representations as a Latent Space for Efficient Generation (Anon et al., 2026) View paper
  - [44] Layton: Latent Consistency Tokenizer for 1024-pixel Image Reconstruction and Generation by 256 Tokens (Xie Qing-song, 2025) View paper
  - Multi-Token Continuous Representations (3 papers)
  - [3] Ming-univision: Joint image understanding and generation with a unified continuous tokenizer (Huang, 2025) View paper
  - [5] Atoken: A unified tokenizer for vision (Lu, 2025) View paper
  - [9] SoftVQ-VAE: Efficient 1-Dimensional Continuous Tokenizer (Hao Chen, 2024) View paper
  - Unified Tokenizers for Multiple Modalities (1 papers)
  - [49] MedITok: A Unified Tokenizer for Medical Image Synthesis and Interpretation (Ma Chenglong, 2025) View paper
- Autoregressive Generation with Continuous Tokens
  - Next-Token Prediction in Continuous Space (6 papers)
  - [2] Autoregressive Image Generation without Vector Quantization (Mingyang Deng, 2024) View paper
  - [7] Fast Autoregressive Models for Continuous Latent Generation (Hang, 2025) View paper
  - [10] Rethinking Discrete Tokens: Treating Them as Conditions for Continuous Autoregressive Image Synthesis (Zheng Peng, 2025) View paper
  - [30] Multimodal Latent Language Modeling with Next-Token Diffusion (Sun Yutao, 2024) View paper
  - [33] Continuous Speculative Decoding for Autoregressive Image Generation (Wang Zili, 2024) View paper
  - [35] NextStep-1: Toward Autoregressive Image Generation with Continuous Tokens at Scale (Han Chunrui, 2025) View paper
  - Frequency and Spectral Autoregressive Methods (1 papers)
  - [40] Frequency Autoregressive Image Generation with Continuous Tokens (Yu Hu, 2025) View paper
  - Multistage and Efficient Autoregressive Architectures (1 papers)
  - [23] E-CAR: Efficient Continuous Autoregressive Image Generation via Multistage Modeling (Yuan, 2024) View paper

- Hybrid and Discrete-Continuous Approaches
  - Hybrid Tokenization with Discrete and Continuous Components (1 papers)
  - [4] HART: Efficient Visual Generation with Hybrid Autoregressive Transformer (Tang, 2024) View paper
  - Flow Matching and Vocabulary Alignment (1 papers)
  - [13] V2Flow: Unifying Visual Tokenization and Large Language Model Vocabularies for Autoregressive Image Generation (Zhang Gui-Wei, 2025) View paper
- Unified Multimodal Generation and Understanding
  - Joint Generation and Understanding with Continuous Tokens (2 papers)
  - [1] Unified autoregressive visual generation and understanding with continuous tokens (Fan Lijie, 2025) View paper
  - [45] Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens (Fan Lijie, 2024) View paper
  - Multimodal Latent Space Alignment (2 papers)
  - [15] OmniBridge: Unified Multimodal Understanding, Generation, and Retrieval via Latent Space Alignment (Xiao Teng, 2025) View paper
  - [47] MergeVQ: A Unified Framework for Visual Generation and Representation with Disentangled Token Merging and Quantization (Li Siyuan, 2025) View paper
- Diffusion-Based Generation in Continuous Latent Spaces
  - Latent Diffusion without Variational Autoencoders (1 papers)
  - [14] SVG-T2I: Scaling Up Text-to-Image Latent Diffusion Model Without Variational Autoencoder (Minglei Shi, 2025) View paper
  - Joint Appearance-Geometry and Multi-Attribute Latent Diffusion (1 papers)
  - [6] Orchid: Image Latent Diffusion for Joint Appearance and Geometry Generation (Krishnan, 2025) View paper
  - Continuous-Scale and Super-Resolution Diffusion (1 papers)
  - [21] Latent Diffusion, Implicit Amplification: Efficient Continuous-Scale Super-Resolution for Remote Sensing Images (Hanlin Wu, 2024) View paper
  - Unified In-Context Generation and Editing (1 papers)
  - [50] FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space (Black Forest Labs, 2025) View paper
- Conditional and Controllable Generation
  - Text and Image Co-Conditioning (1 papers)
  - [22] Unified multi-modal latent diffusion for joint subject and text conditional image generation (Ma, 2023) View paper
  - Continuous Attribute Modulation and Regression Labels (3 papers)
  - [31] Ccgan: Continuous conditional generative adversarial networks for image generation (Xin Ding, 2021) View paper
  - [32] Viewpoint Textual Inversion: Discovering Scene Representations and 3D View Control in 2D Diffusion Models (James Burgess, 2023) View paper
  - [39] Latent age attribute modulation guided continuous aging facial image generation (Jinglei Qu, 2025) View paper
  - Spatial and Structural Control (1 papers)
  - [16] Unigs: Unified representation for image generation and segmentation (Lu Qi, 2024) View paper
  - Multimodal Control Unification (1 papers)
  - [38] UFC-BERT: Unifying multi-modal controls for conditional image synthesis (Zhu Zhang, 2021) View paper
- Domain-Specific and Application-Oriented Generation
  - Medical Image Synthesis (3 papers)
  - [24] Medisyn: A generalist text-guided latent diffusion model for diverse medical image synthesis (Cho, 2024) View paper
  - [29] HC3L-Diff: Hybrid conditional latent diffusion with high frequency enhancement for CBCT-to-CT synthesis (Yin Shi, 2024) View paper
  - [41] High-Fidelity Unified One-to-Many Medical Image Synthesis via Text-Conditioned Latent Diffusion (Youjian Zhang, 2025) View paper
  - 3D and Scene-Level Generation (2 papers)
  - [12] LN3DIFF++: Scalable Latent Neural Fields Diffusion for Speedy 3D Generation. (Lan, 2024) View paper
  - [18] Can3Tok: Canonical 3D Tokenization and Latent Modeling of Scene-Level 3D Gaussians (Gao, 2025) View paper
  - Cross-View Synthesis (1 papers)
  - [17] Exo2EgoSyn: Unlocking Foundation Video Generation Models for Exocentric-to-Egocentric Video Synthesis (Mohammad Mahdi, 2025) View paper
  - Face and Expression Generation (3 papers)
  - [26] AnyFace++: A Unified Framework for Free-Style Text-to-Face Synthesis and Manipulation (Jianxin Sun, 2024) View paper
  - [36] 3D Cartoon Face Generation with Controllable Expressions from a Single GAN Image (Hao Wang, 2022) View paper
  - [37] Towards Open-World Text-Guided Face Image Generation and Manipulation (Xia, 2021) View paper
- Representation Learning and Latent Space Properties
  - Latent Space Unification and Formulation (2 papers)
  - [28] Unifying diffusion models' latent space, with applications to cyclediffusion and guidance (Wu, 2022) View paper
  - [42] Latent Zoning Network: A Unified Principle for Generative Modeling, Representation Learning, and Classification (Lin Zi-nan, 2025) View paper
  - Semantic Discovery and Latent Manipulation (1 papers)
  - [11] Attention Distillation: A Unified Approach to Visual Characteristics Transfer (Zhou Yang, 2025) View paper
- Cross-Modal and Multimodal Synthesis
  - Cross-Modal Image Synthesis (1 papers)
  - [19] Unified Cross-Modal Image Synthesis with Hierarchical Mixture of Product-of-Experts (Reuben Dorent, 2024) View paper
  - Emotional and Affective Synthesis (1 papers)
  - [8] Emosym: A symbiotic framework for unified emotional understanding and generation via latent reasoning (Yijie Zhu, 2025) View paper
- Auxiliary Techniques and Architectural Components
  - Few-Shot and Domain Adaptation (1 papers)
  - [48] Keep and Extent: Unified Knowledge Embedding for Few-Shot Image Generation (Chenghao Xu, 2025) View paper
  - Multimodal Image-to-Image Translation (1 papers)
  - [34] Toward multimodal image-to-image translation (Jun Zhu, 2017) View paper

- Continuous Filter and Neural ODE Approaches (1 papers)
- [20] Image generation using continuous filter atoms (Ze Wang, 2021) View paper
- Arbitrary Resolution and Coordinate-Based Generation (1 papers)
- [27] Generating large images from latent vectors (Ha, 2016) View paper
- Joint Task Learning (1 papers)
- [25] Joint Expression Synthesis and Representation Learning for Facial Expression Recognition (Xi Zhang, 2021) View paper
- Unified Visual-Only Frameworks (1 papers)
- [43] UniModel: A Visual-Only Framework for Unified Multimodal Understanding and Generation (Chi Zhang, 2025) View paper
- Character Consistency and Editing (1 papers)
- [46] ReMix: Towards a Unified View of Consistent Character Generation and Editing (Zhou Benjia, 2025) View paper

Narrative

Core task: efficient image generation from single continuous latent token. The field has evolved around the tension between discrete tokenization—long dominant in autoregressive vision models—and emerging continuous representations that promise greater compactness and expressiveness. The taxonomy reflects this landscape through several major branches: Continuous Token Representation and Tokenization explores methods that encode images into smooth latent codes, often achieving ultra-compact single-token or few-token representations (e.g., Layton[44], Adapting Self-Supervised Latent[0]). Autoregressive Generation with Continuous Tokens investigates how to apply sequential modeling without quantization (Autoregressive Without Quantization[2], HART[4]), while Hybrid and Discrete-Continuous Approaches blend both paradigms (SoftVQ VAE[9], Rethinking Discrete Tokens[10]). Unified Multimodal Generation branches (Unified Autoregressive Vision[1], Ming Univision[3]) extend these ideas to joint text-image modeling, and Diffusion-Based Generation in Continuous Latent Spaces leverages diffusion or flow models in smooth embeddings (LN3DIFF[12], V2Flow[13]). Additional branches address conditional control, domain-specific applications, representation learning properties, cross-modal synthesis, and auxiliary architectural components, collectively mapping the diverse strategies for moving beyond discrete codes.

A particularly active line of work focuses on pushing continuous tokenization to its extreme: single-token or ultra-compact representations that drastically reduce sequence length while preserving reconstruction fidelity. Adapting Self-Supervised Latent[0] sits squarely in this cluster, emphasizing how self-supervised pretraining can be adapted to yield a single continuous token per image. Nearby, Layton[44] also explores ultra-compact continuous embeddings, sharing the goal of minimal latent dimensionality. In contrast, methods like Autoregressive Without Quantization[2] and HART[4] retain longer sequences of continuous tokens to enable autoregressive modeling, trading compactness for the flexibility of next-token prediction. Another contrast emerges with hybrid approaches (SoftVQ VAE[9]) that soften discrete codes rather than eliminating them entirely. The central open question across these branches is whether a single continuous token can capture sufficient detail for high-resolution synthesis, or whether a small handful of tokens offers a better balance between efficiency and expressiveness. Adapting Self-Supervised Latent[0] contributes to this debate by demonstrating that leveraging pretrained representations can make single-token schemes more viable, positioning it as a bridge between representation learning and ultra-compact generation.

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, Huang Zhe, ZHANG Yanhao, Lu, et al. (7 authors total) | **Year/Venue**: 2025 | **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 Single-Token and Ultra-Compact Representations category, focusing on efficient image generation through minimal continuous tokens. They overlap in representing images with extremely few tokens (RepTok uses a single continuous token from SSL encoders, while Layton uses 256 discrete tokens bridged to LDM latent space) and both leverage pre-trained models for efficient generation. The key difference is that RepTok operates with a single continuous semantic token from fine-tuned SSL representations and uses flow matching for generation, whereas Layton uses 256 discrete tokens aligned with pre-trained Latent Diffusion Models and employs autoregressive generation with latent consistency models for reconstruction.

Contributions Analysis

Overall novelty summary. ```json { "paragraphs": ["The paper proposes RepTok, a framework representing images with a single continuous latent token derived from self-supervised vision transformers, paired with a flow-matching decoder. According to the taxonomy, it resides in the 'Single-Token and Ultra-Compact Representations' leaf, which contains only two papers total. This leaf sits under the broader 'Continuous Token Representation and Tokenization' branch, indicating a relatively sparse research direction focused on extreme compactness. The sibling paper in this leaf shares the goal of minimal latent dimensionality, suggesting this is an emerging rather than crowded area of investigation.",

```
"The taxonomy reveals neighboring leaves pursuing multi-token continuous representations (three papers) and unified multimodal tokenizers (one paper
"Among three contributions analyzed, the core RepTok framework using a single SSL token examined only one candidate and found one potentially refuta
"Given the sparse taxonomy leaf (two papers) and limited search scope (thirteen candidates), the work appears to occupy a genuine frontier in ultra-
```

] } ```

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

Contribution 1: Representation Tokenizer (RepTok) framework using single continuous SSL token

Description: The authors propose RepTok, a method that adapts pre-trained self-supervised learning (SSL) encoders by fine-tuning only the semantic class token embedding. This single continuous token is paired with a generative decoder trained via flow matching, enabling faithful image reconstruction and efficient generation while eliminating spatial redundancies inherent in 2D latent spaces.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. A Self-supervised Motion Representation for Portrait Video Generation

URL: View paper

Prior Art Analysis

Self-Supervised Motion Portrait[53] demonstrates prior work that uses self-supervised learning to compress motion into a single continuous 1D token for generative modeling. Both papers employ self-supervised learning paradigms to create compact single-token representations that enable efficient generation. The candidate paper's masked motion encoder compresses motion state into 'a compact and abstract latent motion (1d token)' using self-supervised learning, directly paralleling the original paper's approach of using SSL encoders to create single continuous tokens. This establishes that the concept of using SSL-based single continuous tokens for generative tasks existed before the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers propose using self-supervised learning to compress information into a single continuous 1D token. The candidate demonstrates this approach was already applied to motion representation before the original paper's image representation work. - **Original**: we introducerepresentation tokenizer (reptok), a generative modeling framework that represents an image using a single continuous latent token obtained from self-supervised vision transformers. building on a pre-trained ssl encoder, we fine-tune only the semantic token embedding and pair it with a ge... - **Candidate**: we propose semantic latent motion (semo), a compact and expressive motion representation. first, in the abstraction step, we use a carefully designed masked motion encoder, which leverages a self-supervised learning paradigm to compress the subject's motion state into a compact and abstract latent m...

Evidence 2 - **Rationale**: Both papers emphasize the efficiency and quality benefits of their single-token representations, showing that the concept of using compact single tokens for efficient high-quality generation was already established in the candidate work. - **Original**: our single-token formulation resolves the spatial redundancies of the 2d latent space and significantly reduces training costs. despite its simplicity and efficiency, reptok achieves competitive results on class-conditional imagenet generation - **Candidate**: thanks to the compact and expressive nature of semantic latent motion, our method achieves efficient motion representation and high-quality video generation. user studies demonstrate that our approach surpasses state-of-the-art models with an 81% win rate in realism.

Evidence 3 - **Rationale**: Both papers demonstrate using self-supervised learning to create compact semantic tokens that serve as latent representations for generative decoders, establishing prior art for this architectural approach. - **Original**: we show that self-supervised vision transformers can be used more powerfully than just guiding generative training: with minimal adaptation of the semantic token, their smooth and semantically structured latent spaces can directly act as encoders for generative modeling. by injecting the necessary fi... - **Candidate**: we use a carefully designed masked motion encoder, which leverages a self-supervised learning paradigm to compress the subject's motion state into a compact and abstract latent motion (1d token). second, in the reasoning step, we efficiently generate motion sequences based on the driving audio signa...

Contribution 2: Cosine-similarity regularization loss for preserving SSL latent space geometry

Description: A cosine-similarity alignment term is introduced to constrain the fine-tuned token from deviating too far from its pre-trained SSL representation. This regularization maintains the smooth, semantically structured geometry of the original SSL space, which is beneficial for generative modeling, while still allowing the token to integrate fine-grained reconstruction details.

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. Towards latent masked image modeling for self-supervised visual representation learning

URL: View paper

Brief Assessment

Latent Masked Modeling[55] uses cosine-similarity constraints to prevent patch representations from clustering together during masked reconstruction in latent space, not to preserve pre-trained SSL geometry during fine-tuning for generation as in the original paper.

---

2. Improving Local Latent Fingerprint Representations Under Data Constraints

URL: View paper

Brief Assessment

Local Latent Fingerprint[62] uses cosine similarity in a contrastive learning context (NTXent loss) to maximize similarity between minutiae patches and their augmentations for fingerprint identification. This differs fundamentally from the original paper's use of cosine-similarity regularization to constrain fine-tuned tokens from deviating from pre-trained SSL representations during generative modeling.

---

3. Similarity-based Accent Recognition with Continuous and Discrete Self-supervised Speech Representations

URL: View paper

Brief Assessment

Accent Recognition Representations[54] uses cosine similarity for classification (matching accent embeddings to audio features), not as a regularization loss to preserve latent space geometry during fine-tuning for generation tasks.

---

4. EAGLE: Efficient adaptive geometry-based learning in cross-view understanding

URL: View paper

Brief Assessment

EAGLE[60] uses cosine similarity for cross-view geometric constraints in domain adaptation for semantic segmentation, not for preserving SSL latent space geometry in generative modeling. The technical contexts are fundamentally different.

---

5. A self-supervised contrastive learning approach for latent fingerprint identification

URL: View paper

Brief Assessment

Contrastive Latent Fingerprint[57] focuses on fingerprint identification using contrastive learning and cosine similarity for matching, not on preserving SSL latent space geometry during fine-tuning for generative modeling.

---

6. Manifold-Aware Regularization for Self-Supervised Representation Learning

URL: View paper

Brief Assessment

Manifold-Aware Regularization[61] focuses on theoretical frameworks for geometry preservation in SSL training itself, not on adapting pre-trained SSL representations for generative modeling with cosine-similarity constraints.

---

7. Self-supervised representation of non-standard mechanical parts and fine-tuning method integrating macro process knowledge

URL: View paper

Brief Assessment

Self-Supervised Mechanical Parts[58] focuses on representing non-standard mechanical parts using geometric structure and macro process knowledge, not on preserving SSL latent space geometry for generative modeling.

---

8. Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models

URL: View paper

Brief Assessment

SHMT[63] focuses on makeup transfer using latent diffusion models with a 'decoupling-and-reconstruction' paradigm. The paper does not discuss cosine-similarity regularization for preserving SSL latent space geometry in generative modeling contexts.

---

9. Constrained multiview representation for self-supervised contrastive learning

URL: View paper

Brief Assessment

Constrained Multiview Representation[56] uses cosine similarity for multi-view contrastive learning in medical image segmentation, not for preserving SSL latent space geometry during fine-tuning for generative modeling. The technical contexts and objectives differ fundamentally.

---

10. Stabilize the latent space for image autoregressive modeling: A unified perspective

URL: View paper

Brief Assessment

Stabilize Latent Space[59] focuses on stabilizing latent spaces for autoregressive image modeling through k-means clustering on SSL features, not on cosine-similarity regularization to preserve SSL geometry during fine-tuning for generative tasks.

Contribution 3: Lightweight attention-free pipeline for latent generative modeling

Description: The authors demonstrate that by compressing images into a single token, token-to-token interactions become unnecessary, enabling the use of simple MLP-based architectures such as MLP-Mixer instead of attention mechanisms. This drastically reduces training compute while preserving generation quality, achieving competitive ImageNet generation at a fraction of the cost of transformer-based diffusion baselines.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. UNeXt: MLP-based Rapid Medical Image Segmentation Network

URL: View paper

Brief Assessment

UNeXt[52] focuses on medical image segmentation using MLP-Mixer architectures in an encoder-decoder framework, not on latent generative modeling or diffusion-based image generation. The technical domains and objectives are fundamentally different.

---

2. Back to recurrent processing at the crossroad of transformers and state-space models

URL: View paper

Brief Assessment

Recurrent Processing Crossroad[51] discusses attention-free transformers and compressive context history mechanisms, but does not address single-token compression for latent generative modeling or MLP-based architectures for image generation. The candidate focuses on recurrent processing and state-space models rather than the specific compression-generation pipeline proposed in the original work.

Appendix: Text Similarity Detection

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

References

- [0] Adapting Self-Supervised Representations as a Latent Space for Efficient Generation View paper
- [1] Unified autoregressive visual generation and understanding with continuous tokens View paper
- [2] Autoregressive Image Generation without Vector Quantization View paper
- [3] Ming-univision: Joint image understanding and generation with a unified continuous tokenizer View paper
- [4] HART: Efficient Visual Generation with Hybrid Autoregressive Transformer View paper
- [5] Atoken: A unified tokenizer for vision View paper
- [6] Orchid: Image Latent Diffusion for Joint Appearance and Geometry Generation View paper
- [7] Fast Autoregressive Models for Continuous Latent Generation View paper
- [8] Emosym: A symbiotic framework for unified emotional understanding and generation via latent reasoning View paper
- [9] SoftVQ-VAE: Efficient 1-Dimensional Continuous Tokenizer View paper
- [10] Rethinking Discrete Tokens: Treating Them as Conditions for Continuous Autoregressive Image Synthesis View paper
- [11] Attention Distillation: A Unified Approach to Visual Characteristics Transfer View paper
- [12] LN3DIFF++: Scalable Latent Neural Fields Diffusion for Speedy 3D Generation. View paper
- [13] V2Flow: Unifying Visual Tokenization and Large Language Model Vocabularies for Autoregressive Image Generation View paper
- [14] SVG-T2I: Scaling Up Text-to-Image Latent Diffusion Model Without Variational Autoencoder View paper
- [15] OmniBridge: Unified Multimodal Understanding, Generation, and Retrieval via Latent Space Alignment View paper
- [16] Unigs: Unified representation for image generation and segmentation View paper
- [17] Exo2EgoSyn: Unlocking Foundation Video Generation Models for Exocentric-to-Egocentric Video Synthesis View paper
- [18] Can3Tok: Canonical 3D Tokenization and Latent Modeling of Scene-Level 3D Gaussians View paper
- [19] Unified Cross-Modal Image Synthesis with Hierarchical Mixture of Product-of-Experts View paper
- [20] Image generation using continuous filter atoms View paper
- [21] Latent Diffusion, Implicit Amplification: Efficient Continuous-Scale Super-Resolution for Remote Sensing Images View paper
- [22] Unified multi-modal latent diffusion for joint subject and text conditional image generation View paper
- [23] E-CAR: Efficient Continuous Autoregressive Image Generation via Multistage Modeling View paper
- [24] Medisyn: A generalist text-guided latent diffusion model for diverse medical image synthesis View paper
- [25] Joint Expression Synthesis and Representation Learning for Facial Expression Recognition View paper
- [26] AnyFace++: A Unified Framework for Free-Style Text-to-Face Synthesis and Manipulation View paper

- [27] Generating large images from latent vectors View paper
- [28] Unifying diffusion models' latent space, with applications to cyclediffusion and guidance View paper
- [29] HC3L-Diff: Hybrid conditional latent diffusion with high frequency enhancement for CBCT-to-CT synthesis View paper
- [30] Multimodal Latent Language Modeling with Next-Token Diffusion View paper
- [31] Ccgan: Continuous conditional generative adversarial networks for image generation View paper
- [32] Viewpoint Textual Inversion: Discovering Scene Representations and 3D View Control in 2D Diffusion Models View paper
- [33] Continuous Speculative Decoding for Autoregressive Image Generation View paper
- [34] Toward multimodal image-to-image translation View paper
- [35] NextStep-1: Toward Autoregressive Image Generation with Continuous Tokens at Scale View paper
- [36] 3D Cartoon Face Generation with Controllable Expressions from a Single GAN Image View paper
- [37] Towards Open-World Text-Guided Face Image Generation and Manipulation View paper
- [38] UFC-BERT: Unifying multi-modal controls for conditional image synthesis View paper
- [39] Latent age attribute modulation guided continuous aging facial image generation View paper
- [40] Frequency Autoregressive Image Generation with Continuous Tokens View paper
- [41] High-Fidelity Unified One-to-Many Medical Image Synthesis via Text-Conditioned Latent Diffusion View paper
- [42] Latent Zoning Network: A Unified Principle for Generative Modeling, Representation Learning, and Classification View paper
- [43] UniModel: A Visual-Only Framework for Unified Multimodal Understanding and Generation View paper
- [44] Layton: Latent Consistency Tokenizer for 1024-pixel Image Reconstruction and Generation by 256 Tokens View paper
- [45] Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens View paper
- [46] ReMix: Towards a Unified View of Consistent Character Generation and Editing View paper
- [47] MergeVQ: A Unified Framework for Visual Generation and Representation with Disentangled Token Merging and Quantization View paper
- [48] Keep and Extent: Unified Knowledge Embedding for Few-Shot Image Generation View paper
- [49] MedITok: A Unified Tokenizer for Medical Image Synthesis and Interpretation View paper
- [50] FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space View paper
- [51] Back to recurrent processing at the crossroad of transformers and state-space models View paper
- [52] UNeXt: MLP-based Rapid Medical Image Segmentation Network View paper
- [53] A Self-supervised Motion Representation for Portrait Video Generation View paper
- [54] Similarity-based Accent Recognition with Continuous and Discrete Self-supervised Speech Representations View paper
- [55] Towards latent masked image modeling for self-supervised visual representation learning View paper
- [56] Constrained multiview representation for self-supervised contrastive learning View paper
- [57] A self-supervised contrastive learning approach for latent fingerprint identification View paper
- [58] Self-supervised representation of non-standard mechanical parts and fine-tuning method integrating macro process knowledge View paper
- [59] Stabilize the latent space for image autoregressive modeling: A unified perspective View paper
- [60] EAGLE: Efficient adaptive geometry-based learning in cross-view understanding View paper
- [61] Manifold-Aware Regularization for Self-Supervised Representation Learning View paper
- [62] Improving Local Latent Fingerprint Representations Under Data Constraints View paper
- [63] Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models View paper