Novelty Assessment Report

Paper: Latent Diffusion Model without Variational Autoencoder
PDF URL: https://openreview.net/pdf?id=kdpeJNbFyf
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Recent progress in diffusion-based visual generation has largely relied on latent diffusion models with Variational Autoencoders (VAEs). While effective for high-fidelity synthesis, this VAE+Diffusion paradigm still suffers from limited training and inference efficiency, along with poor transferability to broader vision tasks. These issues stem from a key limitation of VAE latent spaces: the lack of clear semantic separation and strong discriminative structure. Our analysis confirms that these properties are not only crucial for perception and understanding tasks, but also equally essential for the stable and efficient training of latent diffusion models. Motivated by this insight, we introduce **SVG**—a novel latent diffusion model without variational autoencoders, which unleashes **S**elf-supervised representations for **V**isual **G**eneration. SVG constructs a feature space with clear semantic discriminability by leveraging frozen DINO features, while a lightweight residual branch captures fine-grained details for high-fidelity reconstruction. Diffusion models are trained directly on this semantically structured latent space to facilitate more efficient learning. As a result, SVG enables accelerated diffusion training, supports few-step sampling, and improves generative quality. Experimental results further show that SVG preserves the semantic and discriminative capabilities of the underlying self-supervised representations, providing a principled pathway toward task-general, high-quality visual representations.

Core Task Landscape

This paper addresses: **Latent Diffusion Models Using Self-Supervised Visual Representations**
A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

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

- **Self-Supervised Representation Learning for Diffusion Models**
- **Diffusion Features for Discriminative Tasks**
- **Domain-Specific Diffusion Applications**
- **Conditional Generation and Control**
- **Self-Supervised Reconstruction and Restoration**
- **Latent Space Design and Autoencoding**

Complete Taxonomy Tree

- Latent Diffusion Models Using Self-Supervised Visual Representations Survey Taxonomy
- Self-Supervised Representation Learning for Diffusion Models
  - Masked Modeling Approaches in Latent Space (3 papers)
  - [1] Towards Latent Masked Image Modeling for Self-Supervised Visual Representation Learning (Yibing Wei, 2024) View paper
  - [33] Masked Diffusion as Self-supervised Representation Learner (Pan Zixuan, 2023) View paper
  - [34] Prompt-SID: Learning Structural Representation Prompt via Latent Diffusion for Single Image Denoising (Huaqiu Li, 2025) View paper
  - Deconstructing Diffusion for Representation Learning (2 papers)
  - [2] Guided diffusion from self-supervised diffusion features (Hu, 2023) View paper
  - [3] Deconstructing denoising diffusion models for self-supervised learning (Chen Xinlei, 2024) View paper
  - Self-Supervised Pretraining Strategies (4 papers)
  - [4] Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion (Junjiao Tian, 2024) View paper
  - [5] Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer (Zhu, 2024) View paper
  - [8] DiNO-diffusion. Scaling medical diffusion via self-supervised pre-training (Jimenez-Perez Guillermo, 2024) View paper
  - [23] Gen-SIS: Generative Self-augmentation Improves Self-supervised Learning (Belagali, 2024) View paper
  - Unified Representation-Generation Frameworks ★ (5 papers)
  - [0] Latent Diffusion Model without Variational Autoencoder (Anon et al., 2026) View paper
  - [17] SODA: Bottleneck Diffusion Models for Representation Learning (Drew A. Hudson, 2023) View paper
  - [27] USP: Unified Self-Supervised Pretraining for Image Generation and Understanding (Chu, 2025) View paper
  - [29] SVG-T2I: Scaling Up Text-to-Image Latent Diffusion Model Without Variational Autoencoder (Minglei Shi, 2025) View paper
  - [37] Aligning visual foundation encoders to tokenizers for diffusion models (Chen Bowei, 2025) View paper
- Diffusion Features for Discriminative Tasks
  - Semantic Segmentation and Object Discovery (2 papers)
  - [14] DiffusionSeg: Adapting Diffusion Towards Unsupervised Object Discovery (Ma Chaofan, 2023) View paper
  - [48] Unsupervised Discovery of 3D Hierarchical Structure with Generative Diffusion Features (Nurislam Tursynbek, 2023) View paper
  - Pose Estimation and Correspondence (3 papers)

    - [13] Diffusion Features for Zero-Shot 6DoF Object Pose Estimation (Bernd Von Gimborn, 2024) View paper
    - [22] Pose-guided self-training with two-stage clustering for unsupervised landmark discovery (Siddharth Tourani, 2024) View paper
    - [40] Unsupervised semantic correspondence using stable diffusion (Hedlin, 2023) View paper
    - Tracking and Temporal Modeling (1 papers)
    - [18] Diff-tracker: text-to-image diffusion models are unsupervised trackers (ZhengBo Zhang, 2024) View paper
    - Deepfake Detection and Interpretability (3 papers)
    - [9] ConceptExpress: Harnessing diffusion models for single-image unsupervised concept extraction (Shaozhe Hao, 2024) View paper
    - [19] Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation (Hang Li, 2023) View paper
    - [39] DeCLIP: Decoding CLIP Representations for Deepfake Localization (Stefan Smeu, 2024) View paper
- Domain-Specific Diffusion Applications
    - Medical Imaging (2 papers)
    - [44] Unsupervised Anomaly Detection in Medical Images Using Masked Diffusion Model (Hasan Iqbal, 2023) View paper
    - [50] SfMDiffusion: self-supervised monocular depth estimation in endoscopy based on diffusion models. (Yu Li, 2025) View paper
    - Robotics and Embodied AI (4 papers)
    - [7] Crossway diffusion: Improving diffusion-based visuomotor policy via self-supervised learning (Xiang Li, 2024) View paper
    - [20] GWM: Towards Scalable Gaussian World Models for Robotic Manipulation (Lu, 2025) View paper
    - [30] StaMo: Unsupervised Learning of Generalizable Robot Motion from Compact State Representation (Liu, 2025) View paper
    - [41] Invisible Servoing: A Visual Servoing Approach with Return-Conditioned Latent Diffusion (Bishoy Gerges, 2024) View paper
    - Neuroscience and Brain Decoding (2 papers)
    - [11] Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding (Zijiao Chen, 2022) View paper
    - [15] Decoding realistic images from brain activity with contrastive self-supervision and latent diffusion (Sun Jingyuan, 2023) View paper
    - Audio and Multimodal Generation (2 papers)
    - [10] AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining (Haohe Liu, 2023) View paper
    - [31] Self-Supervised Audio-Visual Soundscape Stylization (Li, 2024) View paper
    - Remote Sensing and Specialized Imagery (2 papers)
    - [28] Crossdiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model (Yinghui Xing, 2024) View paper
    - [49] Unsupervised Model-Embedded Two-Stage Diffusion Method for Multispectral and Hyperspectral Image Fusion (Jialin Zhou, 2025) View paper
    - Bioimaging and Cellular Analysis (2 papers)
    - [16] Latent 3d graph diffusion (Y You, 2024) View paper
    - [25] LUMIC: Latent diffUsion for Multiplexed Images of Cells (Albert Hung, 2024) View paper
- Conditional Generation and Control
    - Text-to-Image and Instruction-Guided Editing (1 papers)
    - [21] Instruct-CLIP: Improving Instruction-Guided Image Editing with Automated Data Refinement Using Contrastive Learning (Sra, 2025) View paper
    - Image-Conditioned Synthesis (3 papers)
    - [38] WildVidFit: Video Virtual Try-On in the Wild via Image-Based Controlled Diffusion Models (Zijian He, 2024) View paper
    - [43] Improving virtual try-on with garment-focused diffusion models (WAN Siqi, 2024) View paper
    - [46] MV2MV: Multi-View Image Translation via View-Consistent Diffusion Models (Youcheng Cai, 2024) View paper
    - Multi-Scale and Large-Image Generation (2 papers)
    - [12] ZoomLDM: Latent Diffusion Model for multi-scale image generation (Srikar Yellapragada, 2024) View paper
    - [26] Learned representation-guided diffusion models for large-image generation (Alexandros Graikos, 2024) View paper
- Self-Supervised Reconstruction and Restoration
    - Depth Estimation and 3D Reconstruction (1 papers)
    - [6] Jasmine: Harnessing Diffusion Prior for Self-supervised Depth Estimation (Wang Ji-yuan, 2025) View paper
    - Image Denoising and Enhancement (2 papers)
    - [24] Self-supervised ControlNet with Spatio-Temporal Mamba for Real-world Video Super-resolution (Shi, 2025) View paper
    - [47] LMD: Faster Image Reconstruction with Latent Masking Diffusion (Zhiyuan Ma, 2023) View paper
    - Intrinsic Image Decomposition (1 papers)
    - [32] SAIL: Self-supervised Albedo Estimation from Real Images with a Latent Diffusion Model (Djeghim, 2025) View paper
    - Facial and Biometric Representation (2 papers)
    - [42] SELF-SUPERVISED GAIT RECOGNITION WITH DIFFUSION MODEL PRETRAINING (Anuj Kabra, 2025) View paper
    - [45] A Generative Framework for Self-Supervised Facial Representation Learning (Xing Zhen, 2023) View paper
- Latent Space Design and Autoencoding
    - Pretrained Encoder Integration (1 papers)
    - [36] Unsupervised representation learning from pre-trained diffusion probabilistic models (Zhang Zijian, 2022) View paper
    - Latent Space Stabilization (1 papers)
    - [35] Stabilize the latent space for image autoregressive modeling: A unified perspective (Lidong Bing, 2024) View paper

Narrative

Core task: latent diffusion models using self-supervised visual representations. The field has evolved around several complementary directions. Self-Supervised Representation Learning for Diffusion Models explores how to train or adapt diffusion architectures using self-supervised objectives, often unifying representation extraction with generative modeling. Diffusion Features for Discriminative Tasks investigates repurposing pretrained diffusion features for downstream recognition, segmentation, or correspondence problems. Domain-Specific Diffusion Applications tailors diffusion pipelines to specialized modalities such as audio, 3D geometry, or medical imaging. Conditional Generation and Control focuses on steering generation via text, layout, or other guidance signals, while Self-Supervised Reconstruction and Restoration addresses tasks like denoising and inpainting without paired supervision. Finally, Latent Space Design and Autoencoding examines how to construct and stabilize the latent representations that diffusion models operate on, including alternatives to traditional VAE bottlenecks.

A particularly active line of work seeks to merge representation learning with diffusion training, reducing reliance on separate autoencoder stages. Diffusion without VAE[0] exemplifies this trend by directly learning latent codes through self-supervised diffusion objectives, closely related to efforts like Self-Supervised DiT[5] and SODA[17], which also integrate representation discovery into the generative process. Meanwhile, works such as Aligning Foundation Encoders[37] and USP[27] explore how pretrained vision encoders can be aligned or adapted for diffusion latent spaces, offering a middle ground between end-to-end training and fixed VAE pipelines. Compared to these neighbors, Diffusion without VAE[0] emphasizes eliminating the VAE altogether, whereas Aligning Foundation Encoders[37] retains separate encoder modules but seeks better compatibility with diffusion dynamics. This cluster highlights an ongoing tension between architectural simplicity and the flexibility of modular encoder-decoder designs.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. SODA: Bottleneck Diffusion Models for Representation Learning

Authors: Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, et al. (13 authors total) | **Year/Venue**: 2023 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

We introduce SODA, a self-supervised diffusion model, designed for representation learning. The model incorpo-rates an image encoder, which distills a source view into a compact representation, that, in turn, guides the generation of related novel views. We show that by imposing a tight bottleneck between the encoder and a denoising decoder, and leveraging novel view synthesis as a self-supervised ob-jective, we can turn diffusion models into strong represen-tation learners, capable of capturing...

Relationship Analysis

Both papers belong to the unified representation-generation frameworks category, jointly optimizing representation learning and generative modeling through diffusion-based objectives. They overlap in leveraging self-supervised visual representations (DINO features in the original paper, emergent representations in SODA) to construct semantically structured latent spaces for diffusion models, eliminating or reducing reliance on traditional VAE architectures. The key difference is that the original paper explicitly replaces VAE with frozen DINOv3 features augmented by a residual encoder for generation, while SODA learns representations implicitly through a bottleneck diffusion architecture that simultaneously performs novel view synthesis and representation learning without pre-trained feature extractors.

---

2. USP: Unified Self-Supervised Pretraining for Image Generation and Understanding

Authors: Chu, Xiangxiang, LI Renda, Xiangxiang Chu, Wang Yong, et al. (7 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Recent studies have highlighted the interplay between diffusion models and representation learning. Intermediate representations from diffusion models can be leveraged for downstream visual tasks, while self-supervised vision models can enhance the convergence and generation quality of diffusion models. However, transferring pretrained weights from vision models to diffusion models is challenging due to input mismatches and the use of latent spaces. To address these challenges, we propose Unifie...

Relationship Analysis

Both papers belong to the unified representation-generation frameworks category, jointly optimizing representation learning and generative modeling through shared objectives. They overlap in leveraging self-supervised visual representations (DINO/DINOv3) to improve diffusion model training efficiency and generation quality. However, the original paper (SVG) eliminates VAEs entirely by building diffusion models directly on frozen DINOv3 features augmented with a residual encoder, while the candidate paper (USP) retains the VAE framework and uses masked latent modeling in VAE latent space to initialize diffusion models, representing a more incremental modification to the traditional VAE+Diffusion paradigm.

---

3. SVG-T2I: Scaling Up Text-to-Image Latent Diffusion Model Without Variational Autoencoder

Authors: Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Visual generation grounded in Visual Foundation Model (VFM) representations offers a highly promising unified pathway for integrating visual understanding, perception, and generation. Despite this potential, training large-scale text-to-image diffusion models entirely within the VFM representation space remains largely unexplored. To bridge this gap, we scale the SVG (Self-supervised representations for Visual Generation) framework, proposing SVG-T2I to support high-quality text-to-image synthes...

⚠ Similarity Notice

This paper appears to be a direct extension or variant of the original paper. Both papers share the same core methodology (SVG framework), the same fundamental approach of using self-supervised visual representations (DINOv3) for latent diffusion without VAE, and similar technical contributions. The candidate paper (SVG-T2I) appears to be a scaled-up version focusing specifically on text-to-image synthesis, while the original paper presents the foundational SVG framework for general image generation.

---

4. Aligning visual foundation encoders to tokenizers for diffusion models

Authors: Chen Bowei, Bi, Sai, Bowei Chen, Tan Hao, et al. (21 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

In this work, we propose aligning pretrained visual encoders to serve as tokenizers for latent diffusion models in image generation. Unlike training a variational autoencoder (VAE) from scratch, which primarily emphasizes low-level details, our approach leverages the rich semantic structure of foundation encoders. We introduce a three-stage alignment strategy: (1) freeze the encoder and train an adapter and a decoder to establish a semantic latent space; (2) jointly optimize all components with ...

Relationship Analysis

Both papers belong to the unified representation-generation frameworks category, jointly optimizing representation learning and generative modeling through shared objectives. They overlap in leveraging pretrained visual foundation models (DINOv3/visual encoders) to create semantically structured latent spaces for diffusion models, aiming to improve training efficiency and generation quality. The key difference is that the original paper (SVG) completely replaces VAEs by using frozen DINO features with a lightweight residual encoder, while the candidate paper aligns pretrained visual encoders to existing VAE-based tokenizers through a three-stage training strategy that preserves semantic structure while maintaining the VAE framework.

Contributions Analysis

Overall novelty summary. The paper proposes SVG, a latent diffusion model that replaces VAE encoders with frozen DINO features augmented by a lightweight residual branch for detail capture. It resides in the 'Unified Representation-Generation Frameworks' leaf, which contains five papers exploring joint optimization of representation learning and generative modeling. This leaf sits within the broader 'Self-Supervised Representation Learning for Diffusion Models' branch, indicating a moderately active research direction focused on integrating self-supervised objectives directly into diffusion architectures rather than treating encoding and generation as separate stages.

The taxonomy reveals several neighboring approaches. 'Pretrained Encoder Integration' (one paper) and 'Latent Space Stabilization' (one paper) under 'Latent Space Design and Autoencoding' explore similar themes of leveraging discriminative encoders, but focus on stabilization rather than eliminating VAEs entirely. 'Masked Modeling Approaches in Latent Space' (three papers) combines masked reconstruction with latent diffusion, while 'Self-Supervised Pretraining Strategies' (four papers) emphasizes contrastive or generative pretraining before diffusion training. SVG diverges by directly using frozen DINO features as the primary latent space, bypassing both VAE training and masked modeling paradigms.

Among 30 candidates examined, each of the three contributions shows at least one refutable candidate. For 'SVG: latent diffusion model without variational autoencoders', 10 candidates were examined with 1 appearing to provide overlapping prior work. The same pattern holds for 'Analysis of VAE latent space limitations' and 'Unified feature space for multiple vision tasks', each with 10 candidates examined and 1 refutable match. This suggests that within the limited search scope, some prior work addresses similar architectural choices or latent space critiques, though the majority of examined papers do not directly overlap.

Given the search examined 30 semantically similar papers rather than an exhaustive corpus, the analysis captures immediate neighbors but may miss distant or less-cited precedents. The taxonomy structure shows this is a moderately populated area with clear sibling work in unified frameworks, yet the specific combination of frozen DINO features and VAE elimination appears less common among the examined candidates. The refutable matches indicate incremental positioning relative to existing encoder-free or encoder-aligned diffusion methods, though the precise degree of novelty depends on details not fully captured by top-K semantic retrieval.

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

Contribution 1: SVG: latent diffusion model without variational autoencoders

Description: The authors propose SVG, a new latent diffusion framework that replaces the conventional VAE+Diffusion paradigm by constructing a feature space using frozen DINO features augmented with a lightweight residual branch. This approach enables more efficient diffusion training while preserving semantic discriminability.

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. Crossdiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model

URL: View paper

Brief Assessment

Crossdiff Pansharpening[28] focuses on pansharpening (fusing panchromatic and multispectral satellite images) using a cross-predictive diffusion model with frozen encoders. It does not address general-purpose latent diffusion without VAEs or unified visual representations across diverse vision tasks.

---

2. Denoising diffusion models for anomaly localization in medical images

URL: View paper

Brief Assessment

Anomaly Localization Diffusion[56] focuses on anomaly localization in medical images using diffusion models for reconstruction-based anomaly detection, not on constructing latent diffusion frameworks without VAEs for general visual generation tasks.

---

3. Voice-to-Face Generation: Couple of Self-Supervised Representation Learning with Diffusion Model

URL: View paper

Brief Assessment

Voice-to-Face[52] focuses on cross-modal voice-to-face generation using self-supervised representations for voice-face alignment, not on replacing VAEs in latent diffusion models for general visual generation.

---

4. Diffusion adversarial representation learning for self-supervised vessel segmentation

URL: View paper

Brief Assessment

Vessel Segmentation Diffusion[54] applies diffusion models to vessel segmentation tasks using adversarial learning, not to general visual generation. It does not propose replacing VAE+Diffusion paradigms with self-supervised representations for image synthesis.

---

5. Bitrate-Controlled Diffusion for Disentangling Motion and Content in Video

URL: View paper

Brief Assessment

Bitrate-Controlled Diffusion[57] focuses on disentangling motion and content in video sequences using low-bitrate vector quantization and diffusion models, not on replacing VAEs with self-supervised representations for image generation tasks.

---

6. Automated Learning of Semantic Embedding Representations for Diffusion Models

URL: View paper

Brief Assessment

Semantic Embedding Learning[53] focuses on learning representations through denoising diffusion models with timestep-dependent encoders, not on replacing VAEs with self-supervised features for latent diffusion.

---

7. Diffusion based representation learning

URL: View paper

Brief Assessment

Diffusion Representation Learning[55] focuses on representation learning for downstream tasks using diffusion models with time-conditioned encoders, not on replacing VAEs in latent diffusion frameworks for visual generation. The candidate's encoder learns denoising-relevant features across noise levels, while SVG constructs a semantically discriminative feature space using frozen DINO features for efficient diffusion training and generation.

---

8. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer

URL: View paper

Brief Assessment

Self-Supervised DiT[5] focuses on improving diffusion transformer training through self-supervised discrimination in a teacher-student framework, not on replacing VAEs with self-supervised representations for latent space construction. The candidate uses standard VAE latent spaces (SD-VAE) as shown in their experiments, maintaining the VAE+diffusion paradigm rather than eliminating it.

---

9. Diffusion transformers with representation autoencoders

URL: View paper

Prior Art Analysis

Representation Autoencoders[51] demonstrates that the concept of replacing VAEs with pretrained representation encoders for latent diffusion was already explored prior to the original paper. Both papers propose using frozen pretrained encoders (DINO, SigLIP, MAE) augmented with lightweight components to construct feature spaces for diffusion training, eliminating the need for traditional VAE architectures. The candidate explicitly states this approach as their core contribution, using nearly identical technical components and motivations.

Evidence

Evidence 1 - **Rationale**: Both papers propose the same core innovation: replacing VAEs with pretrained representation encoders (specifically mentioning DINO) for latent diffusion. The candidate explicitly frames this as their contribution, demonstrating prior work exists on this exact concept. - **Original**: we introducesvg-a novel latent diffusion model without variational autoencoders, which unleashes self-supervised representations for visual generation. svg constructs a feature space with clear semantic discriminability by leveraging frozen dino features, while a lightweight residual branch captures... - **Candidate**: in this work, we explore replacing the vae with pretrained representation encoders (e.g., dino, siglip, mae) paired with trained decoders, forming what we term representation autoencoders (raes). these models provide both high-quality reconstructions and semantically rich latent spaces, while allowi...

Evidence 2 - **Rationale**: The candidate identifies the same limitations of VAE-based approaches that motivate the original paper's contribution, suggesting this analysis and the resulting solution were already established in prior work. - **Original**: The svg autoencoder is designed to preserve the semantic structure of frozen dino features while supplementing them with residual perceptual information that is crucial for faithful image reconstruction. concretely, it consists of two components: a frozen dinov3 encoder and a lightweight residual en... - **Candidate**: most dits continue to rely on the original vae encoder, which introduces several limitations: outdated backbones that compromise architectural simplicity, low-dimensional latent spaces that restrict information capacity, and weak representations that result from purely reconstruction-based training ...

---

10. Deconstructing denoising diffusion models for self-supervised learning

URL: View paper

Brief Assessment

Deconstructing Denoising Diffusion[3] focuses on representation learning from denoising diffusion models by deconstructing them into classical denoising autoencoders, not on building latent diffusion frameworks without VAEs for generation tasks.

Contribution 2: Analysis of VAE latent space limitations for diffusion models

Description: The authors systematically analyze mainstream VAE latent spaces and demonstrate that the lack of clear semantic separation and discriminative structure in VAE latents hinders efficient diffusion model training, motivating the need for semantically structured feature spaces.

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. Latent Diffusion Autoencoders: Toward Efficient and Meaningful Unsupervised Representation Learning in Medical Imaging

URL: View paper

Brief Assessment

Latent Diffusion Autoencoders[75] focuses on medical imaging applications with diffusion in compressed latent spaces for computational efficiency, not on analyzing semantic discriminability limitations of VAE latent spaces for diffusion training.

---

2. Enhanced medical image generation through advanced latent space diffusion

URL: View paper

Brief Assessment

Enhanced Medical Generation[71] focuses on medical image generation using VAEs with separable self-attention mechanisms, but does not analyze VAE latent space semantic structure or discriminability limitations for diffusion training.

---

3. DDAE++: Enhancing Diffusion Models Towards Unified Generative and Discriminative Learning

URL: View paper

Brief Assessment

DDAE++[76] focuses on architectural improvements through self-conditioning mechanisms to enhance semantic flow in diffusion models, rather than analyzing VAE latent space limitations or semantic discriminability issues that motivate the original work.

---

4. Automated Learning of Semantic Embedding Representations for Diffusion Models

URL: View paper

Brief Assessment

Semantic Embedding Learning[53] does not analyze VAE latent space limitations or semantic discriminability issues in VAE-based diffusion training.

---

5. Exploring representation-aligned latent space for better generation

URL: View paper

Prior Art Analysis

Representation-Aligned Latent[68] demonstrates that prior work has already identified and addressed the lack of semantic structure in VAE latent spaces for diffusion training. The candidate explicitly states that 'traditional vae latents are often seen as spatial compression in pixel space and lack explicit semantic representations, which are essential for modeling the real world' and proposes a solution (REALS) that 'integrates semantic priors to improve generation performance.' This shows that the limitation of VAE latents lacking

semantic discriminability was recognized and addressed before the original paper's submission, directly challenging the novelty of the original paper's analysis.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same fundamental limitation: VAE latents lack semantic structure. The candidate paper explicitly recognizes this problem and proposes a solution, demonstrating that this analysis was not novel to the original paper. - **Original**: we systematically analyze the limitations of mainstream v ae latent spaces in latent diffusion models, highlighting how semantic dispersion may affect the efficiency of generative modeling. - **Candidate**: traditional vae latents are often seen as spatial compression in pixel space and lack explicit semantic representations, which are essential for modeling the real world.

Evidence 2 - **Rationale**: The candidate paper not only identifies the importance of semantic structure for diffusion training but also provides experimental validation of this insight, showing that addressing this limitation improves generation performance by 15% in FID metric. - **Original**: our analysis confirms that these properties are not only crucial for perception and understanding tasks, but also equally essential for the stable and efficient training of latent diffusion models. - **Candidate**: in this paper, we introduce reals (representation-aligned latent space), which integrates semantic priors to improve generation performance. extensive experiments show that fundamental dit and sit trained on reals can achieve a 15% improvement in fid metric.

Evidence 3 - **Rationale**: Both papers identify that VAE latent quality limitations affect generation performance. The candidate explicitly states that VAE latents lack semantic representations, which parallels the original paper's claim about lack of semantic separation and discriminative structure. - **Original**: while effective for highfidelity synthesis, this v ae+diffusion paradigm still suffers from limited training and inference efficiency, along with poor transferability to broader vision tasks. these issues stem from a key limitation of v ae latent spaces: the lack of clear semantic separation and str... - **Candidate**: while this generative paradigm speeds up training and inference, the quality of the generated outputs is limited by the latents' quality. traditional vae latents are often seen as spatial compression in pixel space and lack explicit semantic representations, which are essential for modeling the real...

---

6. Litevae: Lightweight and efficient variational autoencoders for latent diffusion models

URL: View paper

Brief Assessment

LiteVAE[70] focuses on improving VAE computational efficiency through wavelet transforms rather than analyzing semantic discriminability in VAE latent spaces. The paper does not systematically examine semantic separation or discriminative structure as limitations for diffusion training.

---

7. Contrastive conditional latent diffusion for audio-visual segmentation

URL: View paper

Brief Assessment

Contrastive Audio-Visual[73] focuses on audio-visual segmentation using latent diffusion models for conditional generation tasks, not on analyzing VAE latent space limitations or semantic discriminability issues in diffusion training.

---

8. Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation

URL: View paper

Brief Assessment

SceneFactor[74] focuses on 3D scene generation using semantic and geometric latent spaces for controllable editing, not on analyzing VAE latent space limitations for diffusion training. The paper does not systematically examine VAE semantic discriminability issues.

---

9. Latent diffusion model-enabled low-latency semantic communication in the presence of semantic ambiguities and wireless channel noises

URL: View paper

Brief Assessment

Semantic Communication Diffusion[72] focuses on semantic communication systems with channel noise robustness and outlier handling, not on analyzing VAE latent space semantic structure for diffusion training efficiency.

---

10. Bridging generative and discriminative models for unified visual perception with diffusion priors

URL: View paper

Brief Assessment

Unified Visual Perception[69] focuses on leveraging pre-trained diffusion models for discriminative perception tasks (retrieval, classification, segmentation), not on analyzing VAE latent space properties for diffusion training efficiency.

Contribution 3: Unified feature space for multiple vision tasks

Description: The authors demonstrate that SVG constructs a unified feature space that retains the potential to support diverse core vision tasks beyond generation, including perception and understanding, while simultaneously enabling high-quality visual generation.

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. Multimodal intelligence: Representation learning, information fusion, and applications

URL: View paper

Brief Assessment

Multimodal Intelligence[62] is a review paper focusing on multimodal learning (vision + language) for applications like image captioning and VQA, not on constructing unified feature spaces for generation, perception, and understanding tasks within a single vision modality.

---

2. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning

URL: View paper

Brief Assessment

UniMO[63] focuses on unified-modal understanding and generation via cross-modal contrastive learning between text and images, not on constructing a unified feature space for diverse core vision tasks (perception, understanding, and generation) as described in the original contribution.

---

3. Towards more unified in-context visual understanding

URL: View paper

Brief Assessment

Unified Visual Understanding[67] focuses on in-context learning with multimodal input/output (image-to-image/text tasks like segmentation and captioning), not on constructing a unified feature space that simultaneously supports generation, perception, and understanding as claimed in the original paper.

---

4. Towards unified bijective image-text generation for text-to-image person re-identification

URL: View paper

Brief Assessment

Unified Bijective Generation[64] focuses on text-to-image person re-identification through bijective text-image generation, not on constructing unified feature spaces for diverse vision tasks like generation, perception, and understanding.

---

5. Haploomni: Unified single transformer for multimodal video understanding and generation

URL: View paper

Brief Assessment

HaploOmni[61] focuses on unified multimodal understanding and generation using a single transformer architecture, but does not address the specific concept of constructing a unified feature space that simultaneously supports generation, perception, and understanding tasks through self-supervised representations as described in the original paper.

---

6. Gpt4point: A unified framework for point-language understanding and generation

URL: View paper

Brief Assessment

GPT4Point[60] focuses on point cloud-language understanding and generation tasks, not on constructing a unified feature space for diverse 2D vision tasks (perception, understanding, generation) as claimed in the original paper about visual generation with self-supervised representations.

---

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

URL: View paper

Prior Art Analysis

Tokenflow[59] demonstrates a unified feature space that supports both multimodal understanding and generation tasks simultaneously, predating the ORIGINAL paper's claim. The candidate explicitly addresses the challenge of creating a single representation that works across diverse vision tasks including perception, understanding, and generation. Tokenflow's dual-codebook architecture with shared mapping enables 'joint representation learning at both semantic and pixel level' and the paper states 'our method bridges the gap between generation and understanding tasks. This unified representation enables a single tokenizer to excel in both domains.' The candidate validates this unified capability through extensive experiments showing state-of-the-art performance in both understanding benchmarks and generation tasks, directly challenging the novelty of SVG's claim to construct 'a unified feature space that retains the potential to support diverse core vision tasks beyond generation, including perception and understanding.'

Evidence

Evidence 1 - **Rationale**: Both papers claim to create a unified feature space supporting multiple vision tasks. Tokenflow[59] explicitly demonstrates this capability through its dual-codebook architecture that enables both understanding and generation. - **Original**: svg constructs a unified feature space that retains the potential to support multiple core vision tasks beyond generation - **Candidate**: our method bridges the gap between generation and understanding tasks. This unified representation enables a single tokenizer to excel in both domains.

Evidence 2 - **Rationale**: Both papers claim to provide task-general visual representations that preserve semantic capabilities while enabling generation. Tokenflow[59] achieves this through dual codebooks with shared mapping, demonstrating the concept before the ORIGINAL paper. - **Original**: we demonstrate that svg preserves the semantic and discriminative capabilities of the underlying self-supervised representations, providing a principled pathway toward task-general, high-quality visual representations - **Candidate**: tokenflow addresses this challenge through an innovative dual-codebook architecture that decouples semantic and pixel-level feature learning while maintaining their alignment via a shared mapping mechanism. this design enables direct access to both high-level semantic representations crucial for und...

Evidence 3 - **Rationale**: Both approaches use dual encoders to capture semantic and fine-grained features. Tokenflow[59] uses semantic and pixel encoders with pretrained initialization, similar to SVG's frozen DINO features plus residual branch. - **Original**: svg constructs a feature space with clear semantic discriminability by leveraging frozen dino features, while a lightweight residual branch captures fine-grained details for high-fidelity reconstruction - **Candidate**: we propose a dual-encoder architecture comprising a semantic encoder esem and a pixel encoder epix. this design enables the extraction of two distinct types of image features. for the semantic encoder, we initialize it with a pre-trained text-aligned vision encoder (e.g., clip vit-b/14). this initia...

Evidence 4 - **Rationale**: Both papers validate that their unified feature spaces preserve semantic capabilities while enabling high-quality reconstruction and generation, demonstrating similar contributions. - **Original**: experimental results further show that svg preserves the semantic and discriminative capabilities of the underlying self-supervised representations - **Candidate**: tokenflow demonstrates superior reconstruction quality across all metrics in 384x384 resolution-a standard size in multimodal understanding tasks. these results validate the effectiveness of dual codebook design in preserving fine-grained visual details. moreover, the incorporation of shared mapping...

---

8. Vila-u: a unified foundation model integrating visual understanding and generation

URL: View paper

Brief Assessment

Vila-U[58] focuses on unifying understanding and generation through autoregressive next-token prediction, not on constructing a unified feature space that supports perception, understanding, and generation simultaneously as SVG does.

---

9. Are Unified Vision-Language Models Necessary: Generalization Across Understanding and Generation

URL: View paper

Brief Assessment

Unified Vision-Language[66] focuses on mutual enhancement between understanding and generation in vision-language models through mixed training, not on constructing a unified feature space from self-supervised representations for diverse core vision tasks as in the original paper.

---

10. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing

URL: View paper

Brief Assessment

Vitron[65] focuses on a unified pixel-level vision LLM architecture that integrates multiple vision specialists (encoders and decoders) for different tasks, rather than constructing a single unified feature space that inherently supports diverse tasks through its semantic structure as claimed in the original paper.

Appendix: Text Similarity Detection

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

References

- [0] Latent Diffusion Model without Variational Autoencoder View paper
- [1] Towards Latent Masked Image Modeling for Self-Supervised Visual Representation Learning View paper
- [2] Guided diffusion from self-supervised diffusion features View paper
- [3] Deconstructing denoising diffusion models for self-supervised learning View paper
- [4] Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion View paper
- [5] Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer View paper
- [6] Jasmine: Harnessing Diffusion Prior for Self-supervised Depth Estimation View paper
- [7] Crossway diffusion: Improving diffusion-based visuomotor policy via self-supervised learning View paper
- [8] DiNO-diffusion. Scaling medical diffusion via self-supervised pre-training View paper
- [9] ConceptExpress: Harnessing diffusion models for single-image unsupervised concept extraction View paper
- [10] AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining View paper
- [11] Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding View paper
- [12] ZoomLDM: Latent Diffusion Model for multi-scale image generation View paper
- [13] Diffusion Features for Zero-Shot 6DoF Object Pose Estimation View paper
- [14] DiffusionSeg: Adapting Diffusion Towards Unsupervised Object Discovery View paper
- [15] Decoding realistic images from brain activity with contrastive self-supervision and latent diffusion View paper
- [16] Latent 3d graph diffusion View paper
- [17] SODA: Bottleneck Diffusion Models for Representation Learning View paper
- [18] Diff-tracker: text-to-image diffusion models are unsupervised trackers View paper
- [19] Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation View paper
- [20] GWM: Towards Scalable Gaussian World Models for Robotic Manipulation View paper
- [21] Instruct-CLIP: Improving Instruction-Guided Image Editing with Automated Data Refinement Using Contrastive Learning View paper
- [22] Pose-guided self-training with two-stage clustering for unsupervised landmark discovery View paper
- [23] Gen-SIS: Generative Self-augmentation Improves Self-supervised Learning View paper
- [24] Self-supervised ControlNet with Spatio-Temporal Mamba for Real-world Video Super-resolution View paper
- [25] LUMIC: Latent diffUsion for Multiplexed Images of Cells View paper
- [26] Learned representation-guided diffusion models for large-image generation View paper
- [27] USP: Unified Self-Supervised Pretraining for Image Generation and Understanding View paper
- [28] Crossdiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model View paper
- [29] SVG-T2I: Scaling Up Text-to-Image Latent Diffusion Model Without Variational Autoencoder View paper
- [30] StaMo: Unsupervised Learning of Generalizable Robot Motion from Compact State Representation View paper
- [31] Self-Supervised Audio-Visual Soundscape Stylization View paper
- [32] SAIL: Self-supervised Albedo Estimation from Real Images with a Latent Diffusion Model View paper
- [33] Masked Diffusion as Self-supervised Representation Learner View paper
- [34] Prompt-SID: Learning Structural Representation Prompt via Latent Diffusion for Single Image Denoising View paper
- [35] Stabilize the latent space for image autoregressive modeling: A unified perspective View paper
- [36] Unsupervised representation learning from pre-trained diffusion probabilistic models View paper
- [37] Aligning visual foundation encoders to tokenizers for diffusion models View paper
- [38] WildVidFit: Video Virtual Try-On in the Wild via Image-Based Controlled Diffusion Models View paper
- [39] DeCLIP: Decoding CLIP Representations for Deepfake Localization View paper
- [40] Unsupervised semantic correspondence using stable diffusion View paper
- [41] Invisible Servoing: A Visual Servoing Approach with Return-Conditioned Latent Diffusion View paper
- [42] SELF-SUPERVISED GAIT RECOGNITION WITH DIFFUSION MODEL PRETRAINING View paper
- [43] Improving virtual try-on with garment-focused diffusion models View paper
- [44] Unsupervised Anomaly Detection in Medical Images Using Masked Diffusion Model View paper
- [45] A Generative Framework for Self-Supervised Facial Representation Learning View paper
- [46] MV2MV: Multi-View Image Translation via View-Consistent Diffusion Models View paper
- [47] LMD: Faster Image Reconstruction with Latent Masking Diffusion View paper
- [48] Unsupervised Discovery of 3D Hierarchical Structure with Generative Diffusion Features View paper
- [49] Unsupervised Model-Embedded Two-Stage Diffusion Method for Multispectral and Hyperspectral Image Fusion View paper
- [50] SfMDiffusion: self-supervised monocular depth estimation in endoscopy based on diffusion models. View paper
- [51] Diffusion transformers with representation autoencoders View paper
- [52] Voice-to-Face Generation: Couple of Self-Supervised Representation Learning with Diffusion Model View paper
- [53] Automated Learning of Semantic Embedding Representations for Diffusion Models View paper
- [54] Diffusion adversarial representation learning for self-supervised vessel segmentation View paper
- [55] Diffusion based representation learning View paper
- [56] Denoising diffusion models for anomaly localization in medical images View paper
- [57] Bitrate-Controlled Diffusion for Disentangling Motion and Content in Video View paper
- [58] Vila-u: a unified foundation model integrating visual understanding and generation View paper
- [59] Tokenflow: Unified image tokenizer for multimodal understanding and generation View paper
- [60] Gpt4point: A unified framework for point-language understanding and generation View paper
- [61] Haploomni: Unified single transformer for multimodal video understanding and generation View paper
- [62] Multimodal intelligence: Representation learning, information fusion, and applications View paper

- [63] Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning View paper
- [64] Towards unified bijective image-text generation for text-to-image person re-identification View paper
- [65] Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing View paper
- [66] Are Unified Vision-Language Models Necessary: Generalization Across Understanding and Generation View paper
- [67] Towards more unified in-context visual understanding View paper
- [68] Exploring representation-aligned latent space for better generation View paper
- [69] Bridging generative and discriminative models for unified visual perception with diffusion priors View paper
- [70] Litevae: Lightweight and efficient variational autoencoders for latent diffusion models View paper
- [71] Enhanced medical image generation through advanced latent space diffusion View paper
- [72] Latent diffusion model-enabled low-latency semantic communication in the presence of semantic ambiguities and wireless channel noises View paper
- [73] Contrastive conditional latent diffusion for audio-visual segmentation View paper
- [74] Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation View paper
- [75] Latent Diffusion Autoencoders: Toward Efficient and Meaningful Unsupervised Representation Learning in Medical Imaging View paper
- [76] DDAE++: Enhancing Diffusion Models Towards Unified Generative and Discriminative Learning View paper