Novelty Assessment Report

Paper: What matters for Representation Alignment: Global Information or Spatial Structure?
PDF URL: https://openreview.net/pdf?id=y0UxFtXqXf
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Representation alignment helps generation by distilling representations from a pretrained vision encoder to intermediate diffusion features. We investigate a fundamental question - `what aspect of the target representation matters for generation, its global information (measured by Imagenet1K accuracy) or its spatial structure (pairwise cosine similarity between patch tokens)''? Prevalent wisdom holds that stronger global performance leads to better generation as a target representation. To study this, we first perform a large-scale empirical analysis across 27 different vision encoders and different model scales. The results are surprising - spatial structure, rather than global performance drives the generation performance of a target representation. To further study this, we introduce two straightforward modifications, which specifically accentuate the transfer of spatial information. We replace the standard MLP projection layer in REPA with a simple convolution layer and introduce a spatial normalization layer for the external representation. Surprisingly, our simple method (implemented in <4 lines of code), termed iREPA, consistently improves convergence speed of REPA, across a diverse set of vision encoders, model sizes, and training variants (such as REPA-E and meanflow with REPA). Our work motivates revisiting the fundamental working mechanism of representational alignment and how it can be leveraged for improved training of generative models.

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: **Representation Alignment for Diffusion Model Training**

A total of **50 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

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

- **Core Representation Alignment Methods**
- **Text-Visual Alignment Enhancement**
- **Inference-Time Alignment**
- **Post-Training Optimization**
- **Domain Adaptation and Generalization**
- **Specialized Alignment Applications**
- **Theoretical and Survey Perspectives**

Complete Taxonomy Tree

- Representation Alignment for Diffusion Model Training Survey Taxonomy
- Core Representation Alignment Methods
  - Image Generation Alignment
  - Feature Space Alignment Foundations ★ (3 papers)
    - [0] What matters for Representation Alignment: Global Information or Spatial Structure? (Anon et al., 2026) View paper
    - [1] Representation alignment for generation: Training diffusion transformers is easier than you think (Yu, 2024) View paper
    - [27] Diffusion model as representation learner (Xing-yi Yang, 2023) View paper
  - Spatial Structure Emphasis (2 papers)
    - [15] Aligning visual foundation encoders to tokenizers for diffusion models (Chen Bowei, 2025) View paper
    - [17] Exploring representation-aligned latent space for better generation (Yue, 2025) View paper
  - Multimodal Representation Fusion (2 papers)
    - [30] Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think (Chen Liang, 2025) View paper
    - [44] An Intermediate Fusion ViT Enables Efficient Text-Image Alignment in Diffusion Models (Hu, 2024) View paper
  - Video Generation Alignment
  - Temporal Representation Alignment (3 papers)
    - [2] Cross-frame representation alignment for fine-tuning video diffusion models (Hwang Sung-Won, 2025) View paper
    - [5] Align your latents: High-resolution video synthesis with latent diffusion models (Andreas Blattmann, 2023) View paper
    - [13] CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer (Yang, 2024) View paper
  - Motion-Centric Alignment (2 papers)
    - [7] Moalign: Motion-centric representation alignment for video diffusion models (Bhowmik, 2025) View paper
    - [10] Spectral Motion Alignment for Video Motion Transfer using Diffusion Models (Park, 2024) View paper
  - Specialized Modality Alignment (3 papers)
  - [8] Dual-layer cross-modal alignment recommendation based on the diffusion model (Yuhan Xiu, 2026) View paper
  - [18] Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation (Li Qiulin, 2025) View paper

- [26] MFM-DA: Instance-Aware Adaptor and Hierarchical Alignment for Efficient Domain Adaptation in Medical Foundation Models (Lei, 2025) View paper

- Text-Visual Alignment Enhancement
  - Text Encoding Enhancement (2 papers)
  - [3] Text-image alignment for diffusion-based perception (Kondapaneni, 2024) View paper
  - [22] Ella: Equip diffusion models with llm for enhanced semantic alignment (Hu Xiwei, 2024) View paper
  - Attention-Based Semantic Alignment (2 papers)
  - [47] Attentive Linguistic Tracking in Diffusion Models for Training-free Text-guided Image Editing (Bingyan Liu, 2024) View paper
  - [49] StarVid: Enhancing Semantic Alignment in Video Diffusion Models via Spatial and SynTactic Guided Attention Refocusing (Li Yuanhang, 2024) View paper
  - Long-Text Alignment (1 papers)
  - [9] Improving long-text alignment for text-to-image diffusion models (Liu Luping, 2024) View paper
- Inference-Time Alignment
  - Reward-Guided Generation (2 papers)
  - [19] Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review (Uehara, 2025) View paper
  - [23] Reward-guided controlled generation for inference-time alignment in diffusion models: Tutorial and review (Masatoshi Uehara, 2025) View paper
  - Latent Space Manipulation (3 papers)
  - [14] DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models (Mou, 2023) View paper
  - [21] Aligning optimization trajectories with diffusion models for constrained design generation (Giannone, 2023) View paper
  - [29] Real-world image variation by aligning diffusion inversion chain (Zhang Yue-chen, 2023) View paper
  - Multimodal Guidance Calibration (1 papers)
  - [39] Img: Calibrating diffusion models via implicit multimodal guidance (Guo Jiayi, 2025) View paper
- Post-Training Optimization
  - Preference-Based Alignment (3 papers)
  - [6] Diffusion model alignment using direct preference optimization (Bram Wallace, 2024) View paper
  - [41] Preference-Based Alignment of Discrete Diffusion Models (Wells, 2025) View paper
  - [42] Aligning diffusion models by optimizing human utility (Li Shu-Fan, 2024) View paper
  - Reinforcement Learning Alignment (3 papers)
  - [11] Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards (Zijing Hu, 2025) View paper
  - [20] Aligning text-to-image diffusion models with reward backpropagation (Prabhudesai, 2023) View paper
  - [25] Human-Feedback Efficient Reinforcement Learning for Online Diffusion Model Finetuning (Chen, 2024) View paper
  - Adaptive Personalization (2 papers)
  - [16] APT: Adaptive Personalized Training for Diffusion Models with Limited Data (JungWoo Chae, 2025) View paper
  - [24] I Think, Therefore I Diffuse: Enabling Multimodal In-Context Reasoning in Diffusion Models (Mi, 2025) View paper
- Domain Adaptation and Generalization
  - Cross-Domain Feature Alignment (1 papers)
  - [36] Cross-domain diffusion with progressive alignment for efficient adaptive retrieval (JunYu Luo, 2025) View paper
  - Domain Generalization Enhancement (1 papers)
  - [37] Boosting domain generalized and adaptive detection with diffusion models: Fitness, generalization, and transferability (HE Boyong, 2025) View paper
- Specialized Alignment Applications
  - Conditional Generation Alignment (2 papers)
  - [4] Color Alignment in Diffusion (Ka Chun Shum, 2025) View paper
  - [35] ArbiViewGen: Controllable Arbitrary Viewpoint Camera Data Generation for Autonomous Driving via Stable Diffusion Models (Chen Jingfeng, 2025) View paper
  - Multimodal Reasoning Integration (1 papers)
  - [34] Unifying visual and semantic feature spaces with diffusion models for enhanced cross-modal alignment (Zheng, 2024) View paper
  - Structured Generation Alignment (3 papers)
  - [32] Diffdance: Cascaded human motion diffusion model for dance generation (Qiaosong Qi, 2023) View paper
  - [45] Fine-grained Appearance Transfer with Diffusion Models (Ye, 2023) View paper
  - [46] Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing (Haoyu Zhao, 2024) View paper
  - Compression and Efficiency Alignment (2 papers)
  - [12] Enhanced Distribution Alignment for Post-Training Quantization of Diffusion Models (Liu Xuewen, 2024) View paper
  - [43] Toward Extreme Image Compression With Latent Feature Guidance and Diffusion Prior (Zhiyuan Li, 2024) View paper
  - Cross-Modal Recommendation Alignment (1 papers)
  - [33] Diffcl: A diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations (Qiya Song, 2025) View paper
  - Behavior and Preference Customization (1 papers)
  - [38] Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model (Zibin Dong, 2023) View paper
  - Long-Form Generation Alignment (1 papers)
  - [48] Long-form music generation with latent diffusion (Zach Evans, 2024) View paper
  - Physics-Informed Alignment (1 papers)
  - [50] Physics-Informed Representation Alignment for Sparse Radio-Map Reconstruction (Haozhe Jia, 2025) View paper
  - Anomaly Detection Alignment (1 papers)
  - [40] Aligning Normal Representations in Diffusion Model for Video Anomaly Detection (Chongye Guo, 2025) View paper
- Theoretical and Survey Perspectives (2 papers)
  - [28] Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing (Wu Sihao, 2025) View paper
  - [31] Improving alignment and controllability in GANs and diffusion models (Yumeng, 2025) View paper

Narrative

Core task: representation alignment for diffusion model training. The field has organized itself around several major branches that reflect different stages and modalities of alignment. Core Representation Alignment Methods establish foundational techniques for matching feature spaces during training, often focusing on image generation and cross-modal consistency. Text-Visual Alignment Enhancement addresses the challenge of faithfully translating textual descriptions into visual outputs, with works like Text-image Alignment[3] and Long-text Alignment[9] tackling prompt fidelity at different scales. Inference-Time Alignment and Post-Training Optimization branches explore how to refine alignment after initial training, using techniques such as Direct Preference Optimization[6] and reward-guided generation. Domain Adaptation and Generalization methods extend alignment strategies across different data distributions, while Specialized Alignment Applications target specific modalities like video, audio, or 3D content. Theoretical and Survey Perspectives provide overarching frameworks, as seen in Preference Alignment Survey[28] and related tutorial works.

Within the dense Core Representation Alignment Methods branch, a key tension emerges between global feature matching and spatially-aware alignment strategies. Global or Spatial[0] sits at the intersection of these approaches within the Feature Space Alignment Foundations cluster, exploring how different granularities of alignment affect generation quality. This contrasts with nearby works like Representation Alignment Generation[1], which emphasizes end-to-end learned alignment mechanisms, and Representation Learner[27], which focuses on discovering alignment structures from data. Cross-frame Alignment[2] extends similar principles to temporal consistency in video generation. The original paper's emphasis on spatial versus global trade-offs positions it as addressing a fundamental design choice that ripples through many downstream applications, from text-to-image synthesis to domain transfer tasks.

Related Works in Same Category

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

1. Representation alignment for generation: Training diffusion transformers is easier than you think

Authors: Yu, Sihyun, Sihyun Yu, Sangkyung Kwak, Jang, et al. (18 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Recent studies have shown that the denoising process in (generative) diffusion models can induce meaningful (discriminative) representations inside the model, though the quality of these representations still lags behind those learned through recent self-supervised learning methods. We argue that one main bottleneck in training large-scale diffusion models for generation lies in effectively learning these representations. Moreover, training can be made easier by incorporating high-quality extern...

Relationship Analysis

Both papers belong to the Feature Space Alignment Foundations category, establishing methods to align diffusion model representations with pretrained visual encoders for improved image generation. The candidate paper (REPA) introduces the foundational representation alignment framework that distills pretrained self-supervised visual representations into diffusion transformer features through a regularization technique, demonstrating significant training efficiency improvements. The original paper investigates what aspects of target representations matter most for alignment effectiveness, discovering that spatial structure rather than global semantic information drives generation performance, and proposes modifications (iREPA) to accentuate spatial feature transfer—building upon and refining the alignment principles established in REPA.

---

2. Diffusion model as representation learner

Authors: Xing-yi Yang, Xin-Chao Wang, Xingyi Yang, Xinchao Wang | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Diffusion Probabilistic Models (DPMs) have recently demonstrated impressive results on various generative tasks. Despite its promises, the learned representations of pre-trained DPMs, however, have not been fully understood. In this paper, we conduct an in-depth investigation of the representation power of DPMs, and propose a novel knowledge transfer method that leverages the knowledge acquired by generative DPMs for recognition tasks. Our study begins by examining the feature space of DPMs, rev...

Relationship Analysis

Both papers belong to the Feature Space Alignment Foundations category, focusing on establishing alignment between diffusion features and pretrained encoder representations for image generation. The original paper investigates what aspects of target representations (global vs. spatial structure) matter for representation alignment in diffusion model training, while the candidate paper explores diffusion models themselves as representation learners for recognition tasks through knowledge distillation. The key difference is that the original paper studies how to select and align with external pretrained encoders for generation, whereas the candidate paper extracts and transfers knowledge from pretrained diffusion models to non-generative student networks for recognition.

Contributions Analysis

Overall novelty summary. The paper investigates whether spatial structure or global performance of pretrained vision encoders drives representation alignment effectiveness in diffusion models. It sits within the Feature Space Alignment Foundations leaf, which contains three papers establishing foundational alignment techniques for image generation. This leaf is part of the broader Image Generation Alignment subtopic under Core Representation Alignment Methods. The taxonomy shows this is a moderately populated research direction, with sibling papers exploring complementary aspects of feature space alignment but not directly addressing the spatial-versus-global question posed here.

The taxonomy reveals that neighboring leaves emphasize different alignment dimensions: Spatial Structure Emphasis focuses on preserving local feature correspondence, while Multimodal Representation Fusion integrates multiple modalities. The original paper bridges these directions by empirically demonstrating that spatial structure—not global accuracy—predicts alignment success. The taxonomy's scope_note for Feature Space Alignment Foundations explicitly excludes spatial structure emphasis, suggesting the paper's findings challenge existing categorical boundaries. Related branches like Text-Visual Alignment Enhancement and Inference-Time Alignment address orthogonal concerns (prompt fidelity, post-training guidance) rather than the fundamental encoder property question examined here.

Among 30 candidates examined across three contributions, none clearly refute the paper's claims. The large-scale empirical analysis (10 candidates examined, 0 refutable) appears novel in systematically comparing 27 encoders on spatial versus global metrics. The Spatial Structure Metric contribution (10 candidates, 0 refutable) introduces a predictive measure not found in examined prior work. The iREPA training recipe (10 candidates, 0 refutable) proposes simple architectural modifications—convolution layers and spatial normalization—that accentuate spatial transfer. The limited search scope means these findings reflect novelty within top-30 semantic matches, not exhaustive field coverage.

Based on the limited literature search, the paper appears to address an underexplored question within representation alignment: which encoder properties matter most. The taxonomy structure shows the field has organized around alignment mechanisms and modalities but less around encoder selection principles. The empirical scale (27 encoders) and the simplicity of the proposed modifications (under 4 lines of code) suggest practical contributions, though the analysis cannot confirm whether similar spatial-versus-global comparisons exist in work outside the examined candidates.

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

Contribution 1: Large-scale empirical analysis showing spatial structure drives representation alignment effectiveness

Description: The authors conduct extensive experiments across 27 vision encoders and multiple model scales, demonstrating that spatial self-similarity structure (measured by metrics like LDS) correlates much more strongly with generation performance than global semantic information (measured by ImageNet-1K accuracy). This challenges the prevailing assumption that better global performance leads to better 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. HISPACE: Histological Image Synthesis with Pattern And Content Engine

URL: View paper

Brief Assessment

HISPACE[70] focuses on histological image synthesis using pattern and content engines for digital pathology applications. It does not address representation alignment in diffusion models or analyze spatial versus global semantic information in vision encoders for generation tasks.

---

2. Few shot generative model adaption via relaxed spatial structural alignment

URL: View paper

Brief Assessment

Relaxed Spatial Structural[64] focuses on few-shot generative model adaptation through spatial structural alignment in a different context (adapting pretrained generative models with few examples), not on analyzing what drives representation alignment effectiveness in diffusion model training across diverse vision encoders.

---

3. DiffusePast: Diffusion-based Generative Replay for Class Incremental Semantic Segmentation

URL: View paper

Brief Assessment

DiffusePast[65] focuses on class incremental semantic segmentation using diffusion-based generative replay, not on representation alignment for generation tasks or the relationship between spatial structure and global semantics in vision encoders.

---

4. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching

URL: View paper

Brief Assessment

Tablegpt[66] focuses on table-to-text generation using GPT-2 with table structure reconstruction and content matching. It does not address representation alignment in vision models or spatial structure analysis in diffusion transformers for image generation.

---

5. MV-MambaNet: multiscale and multiview visual question answering network for 3D medical images

URL: View paper

Brief Assessment

MV-MambaNet[69] focuses on 3D medical image visual question answering with multiview and multiscale processing. The candidate does not address representation alignment for generative diffusion models or compare spatial versus global semantic features in generation tasks.

---

6. RecompGPT: Generative Pre-trained Transformers-assisted Human Gaze Pattern Learning and Distribution Modeling for Scene Recomposition

URL: View paper

Brief Assessment

RecompGPT[68] focuses on human gaze pattern learning for scene recomposition tasks, not on representation alignment mechanisms in diffusion models or the relationship between spatial structure and generation performance.

---

7. DocLLM: A layout-aware generative language model for multimodal document understanding

URL: View paper

Brief Assessment

DocLLM[61] focuses on document understanding using bounding box spatial information for layout-aware language modeling, not on analyzing vision encoders or representation alignment for diffusion-based generation tasks.

---

8. Inversion-free image editing with language-guided diffusion models

URL: View paper

Brief Assessment

Inversion-free Editing[62] focuses on eliminating the inversion process in diffusion-based image editing through a special variance schedule (DDCM), not on analyzing spatial versus global information in representation alignment for generation tasks.

---

9. Compositional transformers for scene generation

URL: View paper

Brief Assessment

Compositional Transformers[63] focuses on compositional scene generation using transformers with explicit object-oriented structure for GANs, not on analyzing representation alignment mechanisms in diffusion models or comparing spatial versus global features for generation tasks.

---

10. REGLUE Your Latents with Global and Local Semantics for Entangled Diffusion

URL: View paper

Brief Assessment

REGLUE[67] focuses on joint modeling of VAE latents with VFM semantics through a semantic compressor for diffusion models, not on empirical correlation analysis between spatial structure metrics and generation performance across diverse encoders.

Contribution 2: Spatial Structure Metric (SSM) for predicting representation alignment performance

Description: The authors propose several metrics to quantify spatial self-similarity structure between patch tokens, including LDS (local-distant similarity), which measures how cosine similarity varies with spatial distance. These metrics achieve Pearson correlation above 0.85 with generation FID, far exceeding the 0.26 correlation of linear probing accuracy.

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. Neural pattern similarity reveals the inherent intersection of social categories

URL: View paper

Brief Assessment

Neural Pattern Similarity[58] focuses on neural data analysis and subjective perception using spatial auto-correlation functions, not on predicting generative model performance or representation alignment in diffusion transformers.

---

2. Design and evaluation of algorithms for image retrieval by spatial similarity

URL: View paper

Brief Assessment

Spatial Similarity Retrieval[56] focuses on image database retrieval using spatial-orientation graphs between symbolic objects, not on predicting representation alignment performance in generative models or measuring self-similarity structure of patch token representations.

---

3. Multitask representations in the human cortex transform along a sensory-to-motor hierarchy

URL: View paper

Brief Assessment

Sensory-to-Motor Hierarchy[52] focuses on measuring representational similarity across tasks in human cortex using spatial autocorrelation methods, not on predicting generative model performance through spatial self-similarity metrics between patch tokens.

---

4. SP-GEM: Spatial Pattern-Aware Graph Embedding for Matching Multisource Road Networks

URL: View paper

Brief Assessment

SP-GEM[51] focuses on road network matching using spatial patterns in graph embeddings for geographic data, not on predicting representation alignment performance in generative models or measuring spatial self-similarity in vision encoders.

---

5. A multiscale road matching method based on hierarchical road meshes

URL: View paper

Brief Assessment

Multiscale Road Matching[53] addresses road network matching in geographic information systems using spatial similarity for road geometry alignment. This is fundamentally different from the original paper's SSM, which measures spatial self-similarity structure in vision encoder representations to predict diffusion model generation performance.

---

6. The identification of regional forecasting models using space: time correlation functions

URL: View paper

Brief Assessment

Space-time Correlation[60] focuses on spatial-temporal forecasting models using correlation functions for time-series data in regional analysis. This is fundamentally different from the original paper's SSM, which measures spatial self-similarity structure between patch tokens in vision encoders to predict diffusion model generation performance.

---

7. Global optimisation matching method for multi-representation buildings constrained by road network

URL: View paper

Brief Assessment

Global Optimisation Matching[55] focuses on geospatial entity matching for multi-representation buildings using road network constraints and Hungarian algorithm optimization. It does not address representation alignment in generative models or propose metrics for predicting such performance.

---

8. A survey of measures and methods for matching geospatial vector datasets

URL: View paper

Brief Assessment

Geospatial Vector Survey[59] focuses on matching geospatial vector datasets using similarity measures for data integration and conflation. It does not address representation alignment in diffusion models or metrics for predicting generation performance.

---

9. Matching the building footprints of different vector spatial datasets at a similar scale based on one-class support vector machines

URL: View paper

Brief Assessment

Building Footprints Matching[54] addresses geometric matching of building footprints in spatial datasets using one-class SVMs, not representation alignment in generative models or spatial self-similarity metrics for predicting generation performance.

---

10. A Multi-Scale Residential Areas Matching Method Considering Spatial Neighborhood Features

URL: View paper

Brief Assessment

Multi-Scale Residential[57] addresses geographic entity matching using spatial neighborhood features in map data, not representation alignment in generative models or diffusion transformers. The spatial concepts are fundamentally different domains.

Contribution 3: iREPA: improved training recipe accentuating spatial feature transfer

Description: The authors introduce iREPA, which replaces the standard MLP projection layer with a convolutional layer and adds a spatial normalization layer to enhance spatial feature transfer. This simple modification (implemented in fewer than 4 lines of code) consistently improves convergence speed across diverse encoders, model sizes, and training variants including REPA-E and Meanflow with REPA.

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 cellular traffic prediction method based on diffusion convolutional GRU and multi-head attention mechanism

URL: View paper

Brief Assessment

Cellular Traffic Prediction[75] focuses on cellular traffic prediction using diffusion convolutional GRU for spatial-temporal modeling in network traffic, not on improving spatial feature transfer in diffusion models for image generation.

2. Alleviating Distortion in Image Generation via Multi-Resolution Diffusion Models and Time-Dependent Layer Normalization

URL: View paper

Brief Assessment

Multi-Resolution Time-Dependent[80] focuses on multi-resolution diffusion architectures and time-dependent layer normalization for image generation quality, not on spatial feature transfer mechanisms in representation alignment frameworks like iREPA.

3. Diffusion Augmented Flows: Combining Normalizing Flows and Diffusion Models for Accurate Latent Space Mapping

URL: View paper

Brief Assessment

Diffusion Augmented Flows[76] focuses on combining normalizing flows with diffusion models for latent space mapping using U-Net architectures and CNNs for flow transformations. This is architecturally and methodologically distinct from iREPA's approach of using convolutional projection layers and spatial normalization to enhance spatial feature transfer in representation alignment for diffusion transformers.

4. YOLO-SFT: Road Damage Detection Algorithm Based on Feature Diffusion

URL: View paper

Brief Assessment

YOLO-SFT[73] focuses on road damage detection using convolutional layers in a YOLO architecture for feature diffusion in object detection tasks, not on representation alignment for generative diffusion models or spatial feature transfer between encoders and diffusion transformers.

5. Objective detection of eloquent axonal pathways to minimize postoperative deficits in pediatric epilepsy surgery using diffusion tractography and convolutional neural â□¦

URL: View paper

Brief Assessment

Eloquent Axonal Pathways[79] focuses on medical imaging for epilepsy surgery using CNNs for spatial feature detection in brain pathways, not on improving diffusion model training through representation alignment with convolutional layers and spatial normalization.

6. Spooky Action at a Distance: Normalization Layers Enable Side-Channel Spatial Communication

URL: View paper

Brief Assessment

Spooky Action Distance[78] focuses on normalization layers enabling spatial communication in CNNs for localization tasks, not on improving diffusion model training through convolutional projections and spatial normalization for feature transfer from vision encoders.

7. Deep learning for cerebral vascular occlusion segmentation: a novel ConvNeXtV2 and GRN-integrated U-Net framework for diffusion-weighted imaging

URL: View paper

Brief Assessment

ConvNeXtV2 GRN U-Net[72] focuses on medical image segmentation for cerebral vascular occlusion using ConvNeXtV2 blocks in a U-Net architecture, not on diffusion model training or representation alignment for generative models.

8. A quadruple diffusion convolutional recurrent network for human motion prediction

URL: View paper

Brief Assessment

Quadruple Diffusion Convolutional[77] focuses on human motion prediction using diffusion convolutions on skeletal graphs for spatial-temporal modeling, not on improving spatial feature transfer in diffusion models for image generation through convolutional layers and spatial normalization.

9. AplusN: Progressively Integrating Attention and Normalization in Wavelet Domain for Pose Transfer

URL: View paper

Brief Assessment

AplusN[71] focuses on pose-guided person image generation using wavelet domain processing with attention and normalization for texture transfer in human pose synthesis. This is fundamentally different from iREPA's approach of improving diffusion model training through spatial feature transfer using convolutional projection layers and spatial normalization for general image generation tasks.

10. Semantic diffusion network for semantic segmentation

URL: View paper

Brief Assessment

Semantic Diffusion Network[74] focuses on semantic segmentation tasks using anisotropic diffusion processes with semantic difference convolution for boundary enhancement, not on improving diffusion model training through representation alignment or spatial feature transfer in generative models.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Representation alignment for generation: Training diffusion transformers is easier than you think

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] What matters for Representation Alignment: Global Information or Spatial Structure? View paper
- [1] Representation alignment for generation: Training diffusion transformers is easier than you think View paper
- [2] Cross-frame representation alignment for fine-tuning video diffusion models View paper
- [3] Text-image alignment for diffusion-based perception View paper
- [4] Color Alignment in Diffusion View paper
- [5] Align your latents: High-resolution video synthesis with latent diffusion models View paper
- [6] Diffusion model alignment using direct preference optimization View paper
- [7] Moalign: Motion-centric representation alignment for video diffusion models View paper
- [8] Dual-layer cross-modal alignment recommendation based on the diffusion model View paper
- [9] Improving long-text alignment for text-to-image diffusion models View paper
- [10] Spectral Motion Alignment for Video Motion Transfer using Diffusion Models View paper
- [11] Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards View paper
- [12] Enhanced Distribution Alignment for Post-Training Quantization of Diffusion Models View paper
- [13] CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer View paper
- [14] DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models View paper
- [15] Aligning visual foundation encoders to tokenizers for diffusion models View paper
- [16] APT: Adaptive Personalized Training for Diffusion Models with Limited Data View paper
- [17] Exploring representation-aligned latent space for better generation View paper
- [18] Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation View paper
- [19] Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review View paper
- [20] Aligning text-to-image diffusion models with reward backpropagation View paper
- [21] Aligning optimization trajectories with diffusion models for constrained design generation View paper
- [22] Ella: Equip diffusion models with llm for enhanced semantic alignment View paper
- [23] Reward-guided controlled generation for inference-time alignment in diffusion models: Tutorial and review View paper
- [24] I Think, Therefore I Diffuse: Enabling Multimodal In-Context Reasoning in Diffusion Models View paper
- [25] Human-Feedback Efficient Reinforcement Learning for Online Diffusion Model Finetuning View paper
- [26] MFM-DA: Instance-Aware Adaptor and Hierarchical Alignment for Efficient Domain Adaptation in Medical Foundation Models View paper
- [27] Diffusion model as representation learner View paper
- [28] Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing View paper
- [29] Real-world image variation by aligning diffusion inversion chain View paper
- [30] Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think View paper
- [31] Improving alignment and controllability in GANs and diffusion models View paper
- [32] Diffdance: Cascaded human motion diffusion model for dance generation View paper
- [33] Diffcl: A diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations View paper
- [34] Unifying visual and semantic feature spaces with diffusion models for enhanced cross-modal alignment View paper
- [35] ArbiViewGen: Controllable Arbitrary Viewpoint Camera Data Generation for Autonomous Driving via Stable Diffusion Models View paper
- [36] Cross-domain diffusion with progressive alignment for efficient adaptive retrieval View paper
- [37] Boosting domain generalized and adaptive detection with diffusion models: Fitness, generalization, and transferability View paper
- [38] Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model View paper
- [39] Img: Calibrating diffusion models via implicit multimodal guidance View paper
- [40] Aligning Normal Representations in Diffusion Model for Video Anomaly Detection View paper
- [41] Preference-Based Alignment of Discrete Diffusion Models View paper
- [42] Aligning diffusion models by optimizing human utility View paper
- [43] Toward Extreme Image Compression With Latent Feature Guidance and Diffusion Prior View paper
- [44] An Intermediate Fusion ViT Enables Efficient Text-Image Alignment in Diffusion Models View paper
- [45] Fine-grained Appearance Transfer with Diffusion Models View paper
- [46] Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing View paper
- [47] Attentive Linguistic Tracking in Diffusion Models for Training-free Text-guided Image Editing View paper
- [48] Long-form music generation with latent diffusion View paper
- [49] StarVid: Enhancing Semantic Alignment in Video Diffusion Models via Spatial and SynTactic Guided Attention Refocusing View paper
- [50] Physics-Informed Representation Alignment for Sparse Radio-Map Reconstruction View paper
- [51] SP-GEM: Spatial Pattern-Aware Graph Embedding for Matching Multisource Road Networks View paper
- [52] Multitask representations in the human cortex transform along a sensory-to-motor hierarchy View paper
- [53] A multiscale road matching method based on hierarchical road meshes View paper
- [54] Matching the building footprints of different vector spatial datasets at a similar scale based on one-class support vector machines View paper
- [55] Global optimisation matching method for multi-representation buildings constrained by road network View paper
- [56] Design and evaluation of algorithms for image retrieval by spatial similarity View paper
- [57] A Multi-Scale Residential Areas Matching Method Considering Spatial Neighborhood Features View paper
- [58] Neural pattern similarity reveals the inherent intersection of social categories View paper

- [59] A survey of measures and methods for matching geospatial vector datasets View paper
- [60] The identification of regional forecasting models using space: time correlation functions View paper
- [61] DocLLM: A layout-aware generative language model for multimodal document understanding View paper
- [62] Inversion-free image editing with language-guided diffusion models View paper
- [63] Compositional transformers for scene generation View paper
- [64] Few shot generative model adaption via relaxed spatial structural alignment View paper
- [65] DiffusePast: Diffusion-based Generative Replay for Class Incremental Semantic Segmentation View paper
- [66] Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching View paper
- [67] REGLUE Your Latents with Global and Local Semantics for Entangled Diffusion View paper
- [68] RecompGPT: Generative Pre-trained Transformers-assisted Human Gaze Pattern Learning and Distribution Modeling for Scene Recomposition View paper
- [69] MV-MambaNet: multiscale and multiview visual question answering network for 3D medical images View paper
- [70] HISPACE: Histological Image Synthesis with Pattern And Content Engine View paper
- [71] AplusN: Progressively Integrating Attention and Normalization in Wavelet Domain for Pose Transfer View paper
- [72] Deep learning for cerebral vascular occlusion segmentation: a novel ConvNeXtV2 and GRN-integrated U-Net framework for diffusion-weighted imaging View paper
- [73] YOLO-SFT: Road Damage Detection Algorithm Based on Feature Diffusion View paper
- [74] Semantic diffusion network for semantic segmentation View paper
- [75] A cellular traffic prediction method based on diffusion convolutional GRU and multi-head attention mechanism View paper
- [76] Diffusion Augmented Flows: Combining Normalizing Flows and Diffusion Models for Accurate Latent Space Mapping View paper
- [77] A quadruple diffusion convolutional recurrent network for human motion prediction View paper
- [78] Spooky Action at a Distance: Normalization Layers Enable Side-Channel Spatial Communication View paper
- [79] Objective detection of eloquent axonal pathways to minimize postoperative deficits in pediatric epilepsy surgery using diffusion tractography and convolutional neural â View paper
- [80] Alleviating Distortion in Image Generation via Multi-Resolution Diffusion Models and Time-Dependent Layer Normalization View paper