Novelty Assessment Report

Paper: Reconstruction Alignment Improves Unified Multimodal Models
PDF URL: https://openreview.net/pdf?id=ppQWp8yrm7
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Unified multimodal models (UMMs) unify visual understanding and generation within a single architecture. However, conventional training relies on image–text pairs (or sequences) whose captions are typically sparse and miss fine-grained visual details, even when they use hundreds of words to describe a simple image. We introduce **Reconstruction Alignment (RecA)**, a resource-efficient post-training method that leverages visual understanding encoder embeddings as dense "text prompts," providing rich supervision without captions. Concretely, RecA conditions a UMM on its own visual understanding embeddings and optimizes it to reconstruct the input image with a self-supervised reconstruction loss, thereby realigning understanding and generation. Despite its simplicity, RecA is broadly applicable: across autoregressive, masked-autoregressive, and diffusion-based UMMs, it consistently improves generation and editing fidelity. With only 27 GPU-hours, post-training with RecA substantially improves image generation performance on GenEval (0.73 → 0.90) and DPGBench (80.93 → 88.15), while also boosting editing benchmarks (ImgEdit 3.38 → 3.75, GEdit 6.94 → 7.27). Notably, RecA surpasses much larger open-source models and applies broadly across diverse UMM architectures, establishing it as an efficient and general post-training alignment strategy for UMMs.

Core Task Landscape

This paper addresses: **Aligning Visual Understanding and Generation in Unified Multimodal Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Visual Representation and Tokenization Strategies**
- **Unified Architectural Paradigms**
- **Training Strategies and Alignment Methods**
- **Domain-Specific and Specialized Applications**
- **Evaluation, Analysis, and Benchmarking**
- **Modality Integration and Cross-Modal Learning**
- **Surveys and Taxonomies**

Complete Taxonomy Tree

- Aligning Visual Understanding and Generation in Unified Multimodal Models Survey Taxonomy
- Visual Representation and Tokenization Strategies
  - Discrete Visual Tokenization (3 papers)
  - [23] Show-o: One Single Transformer to Unify Multimodal Understanding and Generation (Xie, 2024) View paper
  - [26] Liquid: Language models are scalable and unified multi-modal generators (Wu, 2024) View paper
  - [36] Vision as a Dialect: Unifying Visual Understanding and Generation via Text-Aligned Representations (Han, 2025) View paper
  - Continuous Visual Tokenization (2 papers)
  - [10] Ming-univision: Joint image understanding and generation with a unified continuous tokenizer (Huang, 2025) View paper
  - [25] Unified autoregressive visual generation and understanding with continuous tokens (Fan Lijie, 2025) View paper
  - Dual or Hybrid Visual Vocabularies (4 papers)
  - [2] Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding (Yang Jiao, 2025) View paper
  - [3] Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies (Song Wei, 2025) View paper
  - [7] Harmonizing visual representations for unified multimodal understanding and generation (Wu, 2025) View paper
  - [33] Toklip: Marry visual tokens to clip for multimodal comprehension and generation (Lin Haokun, 2025) View paper
- Unified Architectural Paradigms
  - Autoregressive Unified Models (6 papers)
  - [1] Metamorph: Multimodal understanding and generation via instruction tuning (Tong, 2025) View paper
  - [9] Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action (Jiasen LÃ¼, 2024) View paper
  - [13] Unified-io: A unified model for vision, language, and multi-modal tasks (Lu, 2022) View paper
  - [16] Vila-u: a unified foundation model integrating visual understanding and generation (Wu Yecheng, 2024) View paper
  - [31] VARGPT: Unified Understanding and Generation in a Visual Autoregressive Multimodal Large Language Model (Zhuang Xianwei, 2025) View paper
  - [47] Skywork UniPic: Unified Autoregressive Modeling for Visual Understanding and Generation (Wang Peiyu, 2025) View paper
  - Diffusion-Based Unified Models (3 papers)
  - [17] Univid: The open-source unified video model (Luo Jiabin, 2025) View paper
  - [20] MMaDA: Multimodal Large Diffusion Language Models (Yang Ling, 2025) View paper
  - [35] LMFusion: Adapting Pretrained Language Models for Multimodal Generation (Shi, 2024) View paper

    - Hybrid Autoregressive-Diffusion Models (2 papers)
    - [11] Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset (Chen, 2025) View paper
    - [48] Show-o2: Improved Native Unified Multimodal Models (Xie, 2025) View paper
    - Decoupled Encoding Architectures (2 papers)
    - [12] Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation (Chengyue Wu, 2024) View paper
    - [32] UniFork: Exploring Modality Alignment for Unified Multimodal Understanding and Generation (Li, 2025) View paper
    - State Space and Linear Complexity Models (1 papers)
    - [8] OmniMamba: Efficient and Unified Multimodal Understanding and Generation via State Space Models (Liao, 2025) View paper
- Training Strategies and Alignment Methods
    - Post-Training Alignment and Refinement ★ (2 papers)
    - [0] Reconstruction Alignment Improves Unified Multimodal Models (Anon et al., 2026) View paper
    - [37] ILLUME+: Illuminating Unified MLLM with Dual Visual Tokenization and Diffusion Refinement (Huang, 2025) View paper
    - Instruction Tuning and Task Adaptation (4 papers)
    - [5] InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model (Dong, 2024) View paper
    - [19] Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation (Lin, 2025) View paper
    - [22] SEED-X: Multimodal Models with Unified Multi-granularity Comprehension and Generation (Ge, 2024) View paper
    - [30] Ming-Lite-Uni: Advancements in Unified Architecture for Natural Multimodal Interaction (Gong Biao, 2025) View paper
    - Multi-Stage and Progressive Training (3 papers)
    - [4] Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning (Gao Can, 2021) View paper
    - [18] Lightbagel: A light-weighted, double fusion framework for unified multimodal understanding and generation (Wang, 2025) View paper
    - [27] Univl: A unified video and language pre-training model for multimodal understanding and generation (Luo, 2020) View paper
    - Reinforcement and Reward-Based Learning (1 papers)
    - [43] Unified Multimodal Chain-of-Thought Reward Model through Reinforcement Fine-Tuning (Wang Yi-bin, 2025) View paper
- Domain-Specific and Specialized Applications
    - Video Understanding and Generation (1 papers)
    - [34] Omni-video: Democratizing unified video understanding and generation (Tan Zhi-yu, 2025) View paper
    - Specialized Visual Domains (4 papers)
    - [6] UniSVG: A Unified Dataset for Vector Graphic Understanding and Generation with Multimodal Large Language Models (Jinke Li, 2025) View paper
    - [39] UniF^{2}ace: A Unified Fine-grained Face Understanding and Generation Model (Li Junzhe, 2025) View paper
    - [49] Unifashion: A unified vision-language model for multimodal fashion retrieval and generation (Zhao, 2024) View paper
    - [50] Thinking with Camera: A Unified Multimodal Model for Camera-Centric Understanding and Generation (Liao Kang, 2025) View paper
    - Task-Specific Unified Frameworks (3 papers)
    - [24] Perceive, Understand and Restore: Real-World Image Super-Resolution with Autoregressive Multimodal Generative Models (Wei Hong-yang, 2025) View paper
    - [28] InstruGen: Automatic Instruction Generation for Vision-and-Language Navigation Via Large Multimodal Models (Yan Yu, 2024) View paper
    - [46] Unipose: A unified multimodal framework for human pose comprehension, generation and editing (Yiheng Li, 2025) View paper
- Evaluation, Analysis, and Benchmarking (3 papers)
    - [29] Are Unified Vision-Language Models Necessary: Generalization Across Understanding and Generation (Zhang Ji-hai, 2025) View paper
    - [40] Realunify: Do unified models truly benefit from unification? a comprehensive benchmark (Shi Yang, 2025) View paper
    - [41] Does Understanding Inform Generation in Unified Multimodal Models? From Analysis to Path Forward (Yuwei Niu, 2025) View paper
- Modality Integration and Cross-Modal Learning
    - Pre-Trained Model Adaptation (2 papers)
    - [42] Unilip: Adapting clip for unified multimodal understanding, generation and editing (Tang, 2025) View paper
    - [44] UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation (Lin Bin, 2025) View paper
    - Cross-Modal Contrastive and Unified Pre-Training (2 papers)
    - [14] Dreamllm: Synergistic multimodal comprehension and creation (Dong, 2023) View paper
    - [38] Kosmos-g: Generating images in context with multimodal large language models (Pan Xi-chen, 2023) View paper
    - Multi-Modality Expansion Beyond Vision-Language (1 papers)
    - [15] Visiongpt: Vision-language understanding agent using generalized multimodal framework (Kelly, 2024) View paper
- Surveys and Taxonomies (2 papers)
    - [21] Multimodal image synthesis and editing: A survey and taxonomy (F Zhan, 2023) View paper
    - [45] CatAID: Category-Guided AI-Generated Image Detection via Vision-Language Model Adaptation (Y Cai, 2025) View paper

Narrative

Core task: aligning visual understanding and generation in unified multimodal models. The field has organized itself around several complementary dimensions. Visual Representation and Tokenization Strategies explore how to encode images and videos into discrete tokens that language models can process, with approaches ranging from vector quantization to learned codebooks. Unified Architectural Paradigms investigate end-to-end designs that handle both perception and generation within a single framework, often building on transformer or diffusion backbones. Training Strategies and Alignment Methods address the challenge of teaching models to seamlessly transition between interpreting and creating visual content, including pre-training recipes, instruction tuning, and post-training refinement techniques. Domain-Specific and Specialized Applications adapt these unified models to particular use cases such as medical imaging, video synthesis, or fashion design. Evaluation, Analysis, and Benchmarking provide the metrics and datasets needed to assess cross-modal performance, while Modality Integration and Cross-Modal Learning focus on bridging text, vision, and other signals. Representative works like Janus[12] and Show O[23] illustrate how different architectural choices lead to distinct trade-offs in generation quality versus understanding accuracy.

A particularly active line of work centers on post-training alignment and refinement, where models that have been pre-trained on large-scale data undergo additional tuning to better harmonize their dual capabilities. Reconstruction Alignment[0] exemplifies this direction

by using reconstruction objectives to tighten the coupling between visual encoders and decoders, ensuring that generated outputs remain faithful to understood inputs. This contrasts with approaches like ILLUME Plus[37], which emphasizes iterative refinement and feedback loops to improve generation fidelity. Meanwhile, works such as Metamorph[1] and Dualtoken[3] explore alternative tokenization schemes that may reduce the need for extensive post-training by designing representations that are inherently more aligned. The central tension across these branches involves balancing the expressiveness of visual tokens with the computational cost of alignment, and deciding whether to unify representations early in the architecture or reconcile them through later-stage training. Reconstruction Alignment[0] sits squarely within the post-training refinement cluster, sharing the goal of tightening understanding-generation coherence but differing from neighbors like ILLUME Plus[37] in its emphasis on direct reconstruction losses rather than multi-stage feedback.

Related Works in Same Category

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

1. ILLUME+: Illuminating Unified MLLM with Dual Visual Tokenization and Diffusion Refinement

Authors: Huang, Runhui, Wang Chun-wei, Runhu Huang, Yang Junwei, et al. (26 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

We present ILLUME+ that leverages dual visual tokenization and a diffusion decoder to improve both deep semantic understanding and high-fidelity image generation. Existing unified models have struggled to simultaneously handle the three fundamental capabilities in a unified model: understanding, generation, and editing. Models like Chameleon and EMU3 utilize VQGAN for image discretization, due to the lack of deep semantic interaction, they lag behind specialist models like LLaVA in visual unders...

Relationship Analysis

Both papers belong to the Post-Training Alignment and Refinement category, focusing on improving unified multimodal models after initial training. While the original paper (RECA) introduces a reconstruction-based post-training method using visual understanding embeddings as dense supervision to realign understanding and generation capabilities, ILLUME+ presents a different approach by employing dual visual tokenization (DualViTok) and a diffusion decoder to simultaneously preserve fine-grained textures and text-aligned semantics. The key distinction is that RECA operates as a lightweight post-training alignment strategy applied to existing UMMs, whereas ILLUME+ proposes a comprehensive architectural redesign with dual tokenizers and progressive training procedures.

Contributions Analysis

Overall novelty summary. The paper introduces Reconstruction Alignment (RecA), a post-training method that uses visual understanding embeddings as dense supervision to realign understanding and generation in unified multimodal models. Within the taxonomy, it resides in the 'Post-Training Alignment and Refinement' leaf, which contains only two papers total. This represents a relatively sparse research direction compared to more crowded areas like 'Autoregressive Unified Models' (six papers) or 'Instruction Tuning and Task Adaptation' (four papers), suggesting the specific focus on post-training reconstruction-based alignment is less explored.

The taxonomy reveals that RecA's parent category 'Training Strategies and Alignment Methods' encompasses four distinct approaches: post-training refinement, instruction tuning, multi-stage training, and reinforcement learning. Neighboring leaves like 'Instruction Tuning and Task Adaptation' focus on task-specific adaptation through instruction formats, while 'Multi-Stage and Progressive Training' emphasizes phased learning recipes. RecA diverges by targeting post-training realignment through self-supervised reconstruction rather than instruction-following or progressive curricula. The taxonomy's scope note explicitly distinguishes post-training methods from pre-training and instruction tuning strategies, positioning RecA as addressing a later-stage alignment challenge.

Across three contributions examined, the literature search analyzed thirty candidate papers total, finding zero clearly refutable instances for any contribution. Specifically, the 'Reconstruction Alignment method' examined ten candidates with none providing overlapping prior work; 'Broad applicability across architectures' similarly found no refutations among ten candidates; and 'Efficient post-training achieving SOTA' showed the same pattern. Given the limited search scope of thirty semantically similar papers rather than an exhaustive survey, these statistics suggest that among closely related work examined, no direct precedents for reconstruction-based post-training alignment were identified, though the search scale leaves room for unexamined literature.

Based on the limited thirty-candidate search and sparse taxonomy positioning, RecA appears to occupy a relatively underexplored niche within post-training alignment methods. The absence of refutable prior work among examined candidates, combined with only one sibling paper in its taxonomy leaf, suggests novelty within the scope analyzed. However, the analysis explicitly covers top-K semantic matches rather than comprehensive field coverage, meaning potential related work in adjacent areas like multi-stage training or continuous tokenization may exist outside the examined set.

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

Contribution 1: Reconstruction Alignment (RECA) post-training method

Description: RECA is a self-supervised post-training approach that conditions unified multimodal models on their own visual understanding embeddings to reconstruct input images, thereby realigning understanding and generation capabilities without requiring text captions.

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. GAIR: Improving Multimodal Geo-Foundation Model with Geo-Aligned Implicit Representations

URL: View paper

Brief Assessment

GAIR[73] focuses on multimodal geo-foundation models using contrastive learning across remote sensing and street view imagery with implicit neural representations for geographic alignment. This is fundamentally different from RECA's self-supervised reconstruction approach for unified multimodal models.

---

2. Harnessing Frozen Unimodal Encoders for Flexible Multimodal Alignment

URL: View paper

Brief Assessment

Frozen Unimodal Encoders[69] focuses on aligning frozen vision and language encoders using lightweight projectors for multimodal tasks, not on self-supervised reconstruction-based post-training for unified multimodal models. The technical approaches and objectives differ fundamentally.

---

3. Self-supervised multimodal learning: A survey

URL: View paper

Brief Assessment

Self Supervised Survey[66] is a broad survey on self-supervised multimodal learning covering general objectives, architectures, and alignment strategies. It does not describe RECA's specific approach of conditioning unified multimodal models on their own visual understanding embeddings for image reconstruction as a post-training method.

---

4. S4-Driver: Scalable Self-Supervised Driving Multimodal Large Language Model with Spatio-Temporal Visual Representation

URL: View paper

Brief Assessment

S4-Driver[68] focuses on autonomous driving motion planning using spatio-temporal visual representations in 3D space, not on multimodal alignment through reconstruction. The technical domains and objectives are fundamentally different.

---

5. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation

URL: View paper

Brief Assessment

ZeroNLG[70] focuses on zero-shot cross-lingual and cross-modal NLG by aligning domains in a shared latent space and using unsupervised multilingual auto-encoders for text reconstruction. RECA conditions unified multimodal models on visual understanding embeddings to reconstruct input images for realigning understanding and generation capabilities. These are fundamentally different approaches: ZeroNLG reconstructs text from text coordinates in latent space for zero-shot translation/captioning, while RECA reconstructs images from visual embeddings to improve multimodal generation fidelity.

---

6. A Self-Supervised Generative Fine-Tune of CLIP for VQA for Visually Impaired

URL: View paper

Brief Assessment

Self Supervised CLIP[72] focuses on fine-tuning CLIP for VQA tasks for visually impaired users using synthetic question-answer pairs, not on aligning understanding and generation capabilities through visual encoder embeddings for image reconstruction in unified multimodal models.

---

7. TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation

URL: View paper

Brief Assessment

TokenFlow[71] focuses on a dual-codebook architecture for unified tokenization across understanding and generation tasks, not on post-training alignment methods using visual embeddings as dense prompts for reconstruction.

---

8. Self-supervised multimodal opinion summarization

URL: View paper

Brief Assessment

Multimodal Opinion[74] focuses on opinion summarization from reviews using multimodal encoders (image, metadata, text) to generate summaries. RECA addresses visual-textual alignment in unified multimodal models for image generation/editing via self-supervised reconstruction. These are fundamentally different tasks and methodologies.

---

9. Separating the â□□Chirpâ□□ from the â□□Chatâ□□: Self-supervised Visual Grounding of Sound and Language

URL: View paper

Brief Assessment

Chirp from Chat[75] focuses on audio-visual grounding through contrastive learning of dense image and audio representations for localization tasks. This is fundamentally different from RECA's self-supervised image reconstruction approach using visual encoder embeddings as dense prompts for unified multimodal models.

---

10. Self-supervised multimodal versatile networks

URL: View paper

Brief Assessment

Self Supervised Versatile[67] focuses on self-supervised multimodal learning from videos using contrastive losses between vision, audio, and text modalities. It does not propose conditioning generation models on visual understanding embeddings for image reconstruction as a post-training alignment method.

Contribution 2: Broad applicability across diverse UMM architectures

Description: The method demonstrates generality by delivering consistent performance improvements across multiple unified multimodal model families with different generation mechanisms, including discrete token prediction, masked autoregressive, and continuous diffusion approaches.

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. MMaDA: Multimodal Large Diffusion Language Models

URL: View paper

Brief Assessment

MMaDA[20] focuses on a unified diffusion architecture for multimodal models, while the original paper demonstrates a post-training method (ReCA) that works across autoregressive, masked-autoregressive, and diffusion-based architectures. These represent different technical approaches to achieving broad applicability.

---

2. Chatvla: Unified multimodal understanding and robot control with vision-language-action model

URL: View paper

Brief Assessment

ChatVLA[62] focuses on unifying multimodal understanding and robot control in vision-language-action models, not on improving generation mechanisms across autoregressive, masked-autoregressive, and diffusion architectures for unified multimodal models as described in the original contribution.

---

3. Unigenx: Unified generation of sequence and structure with autoregressive diffusion

URL: View paper

Brief Assessment

UniGenX[61] focuses on a unified framework combining autoregressive and diffusion mechanisms for sequence and structure generation, rather than demonstrating a post-training method's applicability across existing diverse UMM architectures with different generation paradigms (discrete, masked autoregressive, continuous diffusion).

---

4. Unified autoregressive visual generation and understanding with continuous tokens

URL: View paper

Brief Assessment

Unified Autoregressive[25] presents a single unified framework for joint generation and understanding, whereas the original paper demonstrates a post-training method (RECA) that improves multiple existing UMM architectures with different generation mechanisms (discrete, MAR, diffusion). These are fundamentally different approaches to achieving broad applicability.

---

5. Conditional Panoramic Image Generation via Masked Autoregressive Modeling

URL: View paper

Brief Assessment

Conditional Panoramic[65] focuses on panoramic image generation using masked autoregressive modeling for equirectangular projection images, not on unified multimodal models (UMMs) that combine visual understanding and generation across different architectural paradigms (discrete, masked-autoregressive, diffusion).

---

6. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation

URL: View paper

Brief Assessment

Janus[12] focuses on decoupling visual encoders for understanding vs. generation within a single autoregressive architecture, not on post-training methods that improve multiple generation paradigms (autoregressive, masked-autoregressive, diffusion) as the original paper does.

---

7. Lumos-1: On autoregressive video generation from a unified model perspective

URL: View paper

Brief Assessment

Lumos[63] focuses on autoregressive video generation using LLM architectures with minimal modifications, not on post-training methods that improve multiple UMM families (discrete, masked-autoregressive, diffusion) for image generation and editing tasks.

---

8. Skywork UniPic: Unified Autoregressive Modeling for Visual Understanding and Generation

URL: View paper

Brief Assessment

Skywork UniPic[47] focuses on a single unified architecture with specific design choices (decoupled encoding, MAR encoder, progressive training), rather than demonstrating a method that improves multiple existing UMM families with different generation mechanisms as claimed in the original paper.

---

9. Dual diffusion for unified image generation and understanding

URL: View paper

Brief Assessment

Dual Diffusion[64] focuses exclusively on diffusion-based architectures for unified multimodal modeling, whereas the original paper demonstrates improvements across autoregressive, masked-autoregressive, and diffusion-based UMMs. The candidate does not address masked-autoregressive or discrete token prediction architectures.

---

10. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation

URL: View paper

Brief Assessment

Show-o[23] presents a single unified transformer architecture combining autoregressive and discrete diffusion modeling, rather than a method that improves multiple existing UMM architectures with different generation mechanisms as claimed in the original paper.

Contribution 3: Efficient post-training strategy achieving SOTA performance

Description: RECA achieves state-of-the-art results on image generation and editing benchmarks with minimal computational cost, enabling a 1.5B-parameter model to surpass much larger open-source models without requiring GPT-4o distillation data or reinforcement learning.

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. Emu edit: Precise image editing via recognition and generation tasks

URL: View paper

Brief Assessment

Emu Edit[52] focuses on instruction-based image editing through multi-task training across editing and vision tasks, not on post-training strategies for unified multimodal models. The paper does not address efficient post-training methods or compare computational costs in the manner described by RECA.

---

2. Diffusion adversarial post-training for one-step video generation

URL: View paper

Brief Assessment

Diffusion Adversarial[51] focuses on adversarial post-training for one-step video/image generation in diffusion models, while RECA addresses semantic reconstruction alignment for unified multimodal models. The technical approaches, model architectures, and problem domains are fundamentally different.

---

3. Imgedit: A unified image editing dataset and benchmark

URL: View paper

Brief Assessment

ImgEdit[56] focuses on creating a large-scale image editing dataset and benchmark, not on post-training strategies for unified multimodal models. The candidate addresses data curation and evaluation methodology rather than efficient training techniques for achieving SOTA performance.

---

4. Q-dit: Accurate post-training quantization for diffusion transformers

URL: View paper

Brief Assessment

Q-DiT[54] focuses on post-training quantization for diffusion transformers to reduce computational costs, not on post-training alignment for unified multimodal models. The technical approaches (quantization vs. reconstruction alignment) and target domains (model compression vs. generation quality improvement) are fundamentally different.

---

5. Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step

URL: View paper

Brief Assessment

Generate with CoT[60] focuses on chain-of-thought reasoning strategies (test-time verification, DPO alignment) for autoregressive image generation, not on reconstruction-based post-training methods. The technical approaches are fundamentally different.

---

6. Ptqd: Accurate post-training quantization for diffusion models

URL: View paper

Brief Assessment

PTQD[53] focuses on post-training quantization for diffusion models to reduce computational costs, not on post-training alignment for unified multimodal models. The technical approaches and target domains are fundamentally different.

---

7. Factuality Matters: When Image Generation and Editing Meet Structured Visuals

URL: View paper

Brief Assessment

Factuality Matters[55] focuses on structured visual generation (charts, diagrams, mathematical figures) with executable drawing programs and chain-of-thought reasoning, while RECA addresses general image generation/editing through semantic reconstruction alignment. These are fundamentally different problem domains and methodologies.

---

8. Seedream 4.0: Toward next-generation multimodal image generation

URL: View paper

Brief Assessment

SeeDream[58] focuses on a unified multimodal image generation system with diffusion transformers and VAE optimization, while RECA is a reconstruction-based post-training method using visual understanding embeddings. The technical approaches and methodologies are fundamentally different.

---

9. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting

URL: View paper

Brief Assessment

Imagen Editor[59] focuses on text-guided image inpainting using object masking during training, not on general post-training strategies for unified multimodal models. The technical approaches and problem domains differ fundamentally.

---

10. Towards accurate post-training quantization for diffusion models

URL: View paper

Brief Assessment

Accurate PTQ Diffusion[57] focuses on post-training quantization for diffusion models to reduce computational cost, not on general multimodal model post-training for image generation/editing benchmarks. The technical approaches and problem domains differ fundamentally.

Appendix: Text Similarity Detection

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

References

- [0] Reconstruction Alignment Improves Unified Multimodal Models View paper
- [1] Metamorph: Multimodal understanding and generation via instruction tuning View paper
- [2] Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding View paper
- [3] Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies View paper
- [4] Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning View paper
- [5] InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model View paper
- [6] UniSVG: A Unified Dataset for Vector Graphic Understanding and Generation with Multimodal Large Language Models View paper
- [7] Harmonizing visual representations for unified multimodal understanding and generation View paper
- [8] OmniMamba: Efficient and Unified Multimodal Understanding and Generation via State Space Models View paper
- [9] Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action View paper
- [10] Ming-univision: Joint image understanding and generation with a unified continuous tokenizer View paper
- [11] Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset View paper
- [12] Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation View paper
- [13] Unified-io: A unified model for vision, language, and multi-modal tasks View paper
- [14] Dreamllm: Synergistic multimodal comprehension and creation View paper
- [15] Visiongpt: Vision-language understanding agent using generalized multimodal framework View paper
- [16] Vila-u: a unified foundation model integrating visual understanding and generation View paper
- [17] Univid: The open-source unified video model View paper
- [18] Lightbagel: A light-weighted, double fusion framework for unified multimodal understanding and generation View paper

- [19] Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation View paper
- [20] MMaDA: Multimodal Large Diffusion Language Models View paper
- [21] Multimodal image synthesis and editing: A survey and taxonomy View paper
- [22] SEED-X: Multimodal Models with Unified Multi-granularity Comprehension and Generation View paper
- [23] Show-o: One Single Transformer to Unify Multimodal Understanding and Generation View paper
- [24] Perceive, Understand and Restore: Real-World Image Super-Resolution with Autoregressive Multimodal Generative Models View paper
- [25] Unified autoregressive visual generation and understanding with continuous tokens View paper
- [26] Liquid: Language models are scalable and unified multi-modal generators View paper
- [27] Univl: A unified video and language pre-training model for multimodal understanding and generation View paper
- [28] InstruGen: Automatic Instruction Generation for Vision-and-Language Navigation Via Large Multimodal Models View paper
- [29] Are Unified Vision-Language Models Necessary: Generalization Across Understanding and Generation View paper
- [30] Ming-Lite-Uni: Advancements in Unified Architecture for Natural Multimodal Interaction View paper
- [31] VARGPT: Unified Understanding and Generation in a Visual Autoregressive Multimodal Large Language Model View paper
- [32] UniFork: Exploring Modality Alignment for Unified Multimodal Understanding and Generation View paper
- [33] Toklip: Marry visual tokens to clip for multimodal comprehension and generation View paper
- [34] Omni-video: Democratizing unified video understanding and generation View paper
- [35] LMFusion: Adapting Pretrained Language Models for Multimodal Generation View paper
- [36] Vision as a Dialect: Unifying Visual Understanding and Generation via Text-Aligned Representations View paper
- [37] ILLUME+: Illuminating Unified MLLM with Dual Visual Tokenization and Diffusion Refinement View paper
- [38] Kosmos-g: Generating images in context with multimodal large language models View paper
- [39] UniF^{2}ace: A Unified Fine-grained Face Understanding and Generation Model View paper
- [40] Realunify: Do unified models truly benefit from unification? a comprehensive benchmark View paper
- [41] Does Understanding Inform Generation in Unified Multimodal Models? From Analysis to Path Forward View paper
- [42] Unilip: Adapting clip for unified multimodal understanding, generation and editing View paper
- [43] Unified Multimodal Chain-of-Thought Reward Model through Reinforcement Fine-Tuning View paper
- [44] UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation View paper
- [45] CatAID: Category-Guided AI-Generated Image Detection via Vision-Language Model Adaptation View paper
- [46] Unipose: A unified multimodal framework for human pose comprehension, generation and editing View paper
- [47] Skywork UniPic: Unified Autoregressive Modeling for Visual Understanding and Generation View paper
- [48] Show-o2: Improved Native Unified Multimodal Models View paper
- [49] Unifashion: A unified vision-language model for multimodal fashion retrieval and generation View paper
- [50] Thinking with Camera: A Unified Multimodal Model for Camera-Centric Understanding and Generation View paper
- [51] Diffusion adversarial post-training for one-step video generation View paper
- [52] Emu edit: Precise image editing via recognition and generation tasks View paper
- [53] Ptqd: Accurate post-training quantization for diffusion models View paper
- [54] Q-dit: Accurate post-training quantization for diffusion transformers View paper
- [55] Factuality Matters: When Image Generation and Editing Meet Structured Visuals View paper
- [56] Imgedit: A unified image editing dataset and benchmark View paper
- [57] Towards accurate post-training quantization for diffusion models View paper
- [58] Seedream 4.0: Toward next-generation multimodal image generation View paper
- [59] Imagen editor and editbench: Advancing and evaluating text-guided image inpainting View paper
- [60] Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step View paper
- [61] Unigenx: Unified generation of sequence and structure with autoregressive diffusion View paper
- [62] Chatvla: Unified multimodal understanding and robot control with vision-language-action model View paper
- [63] Lumos-1: On autoregressive video generation from a unified model perspective View paper
- [64] Dual diffusion for unified image generation and understanding View paper
- [65] Conditional Panoramic Image Generation via Masked Autoregressive Modeling View paper
- [66] Self-supervised multimodal learning: A survey View paper
- [67] Self-supervised multimodal versatile networks View paper
- [68] S4-Driver: Scalable Self-Supervised Driving Multimodal Large Language Model with Spatio-Temporal Visual Representation View paper
- [69] Harnessing Frozen Unimodal Encoders for Flexible Multimodal Alignment View paper
- [70] Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation View paper
- [71] TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation View paper
- [72] A Self-Supervised Generative Fine-Tune of CLIP for VQA for Visually Impaired View paper
- [73] GAIR: Improving Multimodal Geo-Foundation Model with Geo-Aligned Implicit Representations View paper
- [74] Self-supervised multimodal opinion summarization View paper
- [75] Separating the â□□Chirpâ□□ from the â□□Chatâ□□: Self-supervised Visual Grounding of Sound and Language View paper