Novelty Assessment Report

Paper: UniF^{2}ace: A $\underline{Uni}$fied $\underline{F}$ine-grained $\underline{Face}$ Understanding and Generation Model

PDF URL: https://openreview.net/pdf?id=LV01JdxARe

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Unified multimodal models (UMMs) have emerged as a powerful paradigm in fundamental cross-modality research, demonstrating significant potential in both image understanding and generation. However, existing research in the face domain primarily faces two challenges: **(1) fragmentation development**, with existing methods failing to unify understanding and generation into a single one, hindering the way to artificial general intelligence. **(2) lack of fine-grained facial attributes**, which are crucial for high-fidelity applications. To handle those issues, we propose UniF^{2}ace, the first UMM specifically tailored for fine-grained face understanding and generation. **First**, we introduce a novel theoretical framework with a Dual Discrete Diffusion (D3Diff) loss, unifying masked generative models with discrete score matching diffusion and leading to a more precise approximation of the negative log-likelihood. Moreover, this D3Diff significantly enhances the model's ability to synthesize high-fidelity facial details aligned with text input. **Second**, we propose a multi-level grouped Mixture-of-Experts architecture, adaptively incorporating the semantic and identity facial embeddings to complement the attribute forgotten phenomenon in representation evolvement. **Finally**, to this end, we construct UniF^{2}aceD-1M, a large-scale dataset comprising 130K fine-grained image-caption pairs and 1M visual question-answering pairs, spanning a much wider range of facial attributes than existing datasets. Extensive experiments demonstrate that UniF^{2}ace outperforms existing models with a similar scale in both understanding and generation tasks, with 7.1% higher Desc-GPT and 6.6% higher VQA-score, respectively. Code is available in the supplementary materials.

Disclaimer

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

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

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

Core Task Landscape

This paper addresses: **Unified Fine-Grained Face Understanding and Generation**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Unified Multimodal Face Models**
- **Face Generation and Synthesis**
- **Face Analysis and Recognition**
- **Cross-Resolution Face Enhancement**
- **3D Face Modeling and Reconstruction**
- **Face Modeling Foundations and Surveys**
- **Face Datasets and Benchmarks**

Complete Taxonomy Tree

- Unified Fine-Grained Face Understanding and Generation Survey Taxonomy
- Unified Multimodal Face Models
  - Joint Understanding-Generation Architectures ★ (5 papers)
  - [0] UniF^{2}ace: A $\underline{Uni}$fied $\underline{F}$ine-grained $\underline{Face}$ Understanding and Generation Model (Anon et al., 2026) View paper
  - [1] Uniace: Fine-grained Face Understanding and Generation with Unified Multimodal Models (J Li, 2025) View paper
  - [2] UniCTokens: Boosting Personalized Understanding and Generation via Unified Concept Tokens (Yang Si-han, 2025) View paper
  - [18] Talk2face: A unified sequence-based framework for diverse face generation and analysis tasks (Yudong Li, 2022) View paper
  - [25] A unified deep model for joint facial expression recognition, face synthesis, and face alignment (Feifei Zhang, 2020) View paper
  - General Visual Understanding-Generation Unification (3 papers)
  - [16] Pisces: An Auto-regressive Foundation Model for Image Understanding and Generation (Xu Zhiyang, 2025) View paper
  - [19] Harmonizing Visual Representations for Unified Multimodal Understanding and Generation (Wu, 2025) View paper
  - [27] Unified Autoregressive Visual Generation and Understanding with Continuous Tokens (Fan Lijie, 2025) View paper
- Face Generation and Synthesis
  - Talking Face and Expression Generation (4 papers)
  - [4] Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset (Zhimeng Zhang, 2021) View paper
  - [11] FG-EmoTalk: Talking head video generation with fine-grained controllable facial expressions (Zhaoxu Sun, 2024) View paper
  - [24] Fine-grained talking face generation with video reinterpretation (Xin Huang, 2021) View paper
  - [26] Toward Fine-Grained Talking Face Generation (Zhicheng Sheng, 2023) View paper
  - Pose-Invariant Face Synthesis (3 papers)
  - [20] Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis (Rui Huang, 2017) View paper
  - [21] Learning a high fidelity pose invariant model for high-resolution face frontalization (Cao Jie, 2018) View paper
  - [34] Dual-agent gans for photorealistic and identity preserving profile face synthesis (Jian Zhao, 2017) View paper

  - Identity-Preserving Face Manipulation (3 papers)
  - [31] End-to-end Face-swapping via Adaptive Latent Representation Learning (Lin, 2023) View paper
  - [39] 3D Guided Fine-Grained Face Manipulation (Zhenglin Geng, 2022) View paper
  - [47] Face Adapter for Pre-Trained Diffusion Models with Fine-Grained ID and Attribute Control (Han Yue, 2024) View paper
  - Synthetic Face Data Generation (4 papers)
  - [13] Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model (F Boutros, 2023) View paper
  - [15] More real than real: A study on human visual perception of synthetic faces (Lago, 2021) View paper
  - [38] Privacy-preserving annotation of face images through attribute-preserving face synthesis (Sola Shirai, 2019) View paper
  - [41] Training deep face recognition systems with synthetic data (Kortylewski, 2018) View paper
  - Expression Synthesis and Emotion Modeling (2 papers)
  - [46] SynExpression: A Diffusion-Based Framework for Controllable Facial Expression Synthesis and Emotion Detection Using Facial Segmentation Pose Maps (Shahrzad Sayyafzadeh, 2025) View paper
  - [50] Joint Deep Learning of Facial Expression Synthesis and Recognition (YAN Yan, 2022) View paper
- Face Analysis and Recognition
  - Cross-Spectral and Heterogeneous Face Recognition (4 papers)
  - [14] Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis (Decheng Liu, 2021) View paper
  - [17] Adversarial cross-spectral face completion for NIR-VIS face recognition (Ran He, 2019) View paper
  - [42] Cross-spectral face completion for nir-vis heterogeneous face recognition (He, 2019) View paper
  - [45] ANYRES: Generating High-Resolution visible-face images from Low-Resolution thermal-face images (David Anghelone, 2023) View paper
  - Presentation Attack Detection (1 papers)
  - [12] Presentation attack detection methods for face recognition systems: A comprehensive survey (R Ramachandra, 2017) View paper
  - Multimodal Hallucination Detection (1 papers)
  - [49] Fine-grained Structural Hallucination Detection for Unified Visual Comprehension and Generation in Multimodal LLM (Hao Fei, 2024) View paper
- Cross-Resolution Face Enhancement
  - Face Super-Resolution with Priors (4 papers)
  - [5] A novel generative adversarial networkâ□□based superâ□□resolution approach for face recognition (Amit Chougule, 2024) View paper
  - [8] The face image super-resolution algorithm based on combined representation learning (Yuantao Chen, 2021) View paper
  - [9] Super-resolving very low-resolution face images with supplementary attributes (Xin Yu, 2018) View paper
  - [48] Face super resolution with a high frequency highway (Dan Zeng, 2024) View paper
  - Identity-Aware Face Hallucination (3 papers)
  - [28] Super-Identity Convolutional Neural Network for Face Hallucination (Kaipeng Zhang, 2018) View paper
  - [29] Hallucinating face by eigentransformation (Xiaogang Wang, 2005) View paper
  - [40] Identity aware synthesis for cross resolution face recognition (Maneet Singh, 2018) View paper
  - GAN-Based Face Image Enhancement (1 papers)
  - [32] Facial image synthesis and super-resolution with stacked generative adversarial network (Jijun He, 2020) View paper
- 3D Face Modeling and Reconstruction
  - High-Resolution 3D Face Capture (3 papers)
  - [7] Spacetime faces: high resolution capture for modeling and animation (Li Zhang, 2004) View paper
  - [10] Use of a high resolution 3D optical scanner for 3D model creation, game design and facial expression recognition (Georgia Constantinou, 2023) View paper
  - [43] 3D face recognition based on high-resolution 3D face modeling from frontal and profile views (Lijun Yin, 2003) View paper
  - Monocular 3D Face Reconstruction (2 papers)
  - [30] 3DFaceFill: An analysis-by-synthesis approach to face completion (Rahul Dey, 2022) View paper
  - [36] 3D Facial Expressions through Analysis-by-Neural-Synthesis (George Retsinas, 2024) View paper
  - Few-Shot Volumetric Face Modeling (1 papers)
  - [22] Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures (Marcel C. Buehler, 2024) View paper
  - 3D Face Analysis and Animation (2 papers)
  - [35] Analysis and synthesis of facial image sequences using physical and anatomical models (Demetri Terzopoulos, 2002) View paper
  - [37] Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters (Yongmian Zhang, 2008) View paper
- Face Modeling Foundations and Surveys (4 papers)
  - [3] 3D Face Modelling, Analysis and Synthesis (Koujan, 2022) View paper
  - [6] Face analysis and synthesis (S Morishima, 2001) View paper
  - [23] Facial analysis and synthesis scheme (Ravyse, 2006) View paper
  - [44] Facial analysis and synthesis using image-based models (Tony Ezzat, 1996) View paper
- Face Datasets and Benchmarks (1 papers)
  - [33] Efhq: Multi-purpose extremepose-face-hq dataset (Trung Tuan Dao, 2024) View paper

Narrative

Core task: unified fine-grained face understanding and generation. The field has evolved from early separate pipelines for analysis and synthesis into increasingly integrated architectures that handle multiple face-related tasks within a single framework. The taxonomy reflects this evolution through seven main branches: Unified Multimodal Face Models emphasize joint architectures that combine understanding (e.g., attribute recognition, expression analysis) with generation capabilities; Face Generation and Synthesis focuses on creating realistic faces through GANs, diffusion models, and controllable synthesis methods; Face Analysis and Recognition addresses identity verification, attribute prediction, and anti-spoofing; Cross-Resolution Face Enhancement tackles super-resolution and quality improvement; 3D Face Modeling and Reconstruction builds geometric representations from images; Face Modeling Foundations and Surveys provides theoretical grounding; and Face Datasets and Benchmarks establishes evaluation standards. Works like Uniace[1] and UniCTokens[2] exemplify the push toward unified representations that bridge perception and generation tasks.

Recent developments reveal tension between specialized depth and unified breadth. Some lines pursue task-specific excellence—for instance, super-resolution methods like GAN Super-Resolution[5] or talking face generation approaches such as Flow-guided Talking Face[4] and FG-EmoTalk[11]—while others seek comprehensive frameworks that handle diverse face manipulations simultaneously. UniFace[0] sits squarely within the Joint Understanding-Generation Architectures cluster, sharing conceptual ground with Uniace[1] and Talk2face[18] by attempting to unify perception and synthesis under a single model. Compared to Unified Deep Model[25], which pioneered multi-task face processing, UniFace[0] emphasizes finer-grained control over both semantic understanding and generative quality. The central open question remains whether such unified models can match or exceed specialized systems across all subtasks, or whether hybrid architectures that selectively integrate components will prove more practical for real-world deployment.

Related Works in Same Category

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

1. Uniace: Fine-grained Face Understanding and Generation with Unified Multimodal Models

Authors: J Li, X Qiu, L Xu, L Guo, D Qu, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ containing 130K fine-grained image-caption pairs and one million VQAs. We develop an automated pipeline for generating fine-grained multimodal datasets, leveraging face attribute â¦

⚠ Similarity Notice

These papers appear to be the same work or very closely related variants. Both present UniF2ace as the first unified multimodal model for fine-grained face understanding and generation, introduce the identical Dual Discrete Diffusion (D3Diff) loss and multi-level grouped Mixture-of-Experts architecture, and construct the same UniF2aceD-1M dataset with 130K image-caption pairs and 1M VQA pairs. The titles, abstracts, technical contributions, and experimental results are nearly identical, suggesting these are likely different versions of the same paper.

2. UniCTokens: Boosting Personalized Understanding and Generation via Unified Concept Tokens

Authors: Yang Si-han, Ruichuan An, Zhang, Renrui, Sihan Yang, et al. (27 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Personalized models have demonstrated remarkable success in understanding and generating concepts provided by users. However, existing methods use separate concept tokens for understanding and generation, treating these tasks in isolation. This may result in limitations for generating images with complex prompts. For example, given the concept $\langle bo\rangle$, generating"$\langle bo\rangle$ wearing its hat"without additional textual descriptions of its hat. We call this kind of generation \t...

Relationship Analysis

Both papers belong to the Joint Understanding-Generation Architectures category, developing unified models that perform face analysis and synthesis through shared representations. They overlap in addressing the challenge of unifying understanding and generation tasks within a single framework, with both proposing novel training strategies and architectural innovations (UniF2ace's D3Diff loss and multi-level MoE vs. UniCTokens' unified concept tokens and progressive training). The key difference is that UniF2ace focuses specifically on fine-grained facial attributes with a specialized dataset (UniF2aceD-1M) and domain-specific architectural components, while UniCTokens addresses broader personalized concept understanding and generation across multiple object categories (people, pets, objects) using a unified token-based approach.

3. Talk2face: A unified sequence-based framework for diverse face generation and analysis tasks

Authors: Yudong Li, Xianxu Hou, Zhe Zhao, Linlin Shen, Xuefeng Yang, et al. (6 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Facial analysis is an important domain in computer vision and has received extensive research attention. For numerous downstream tasks with different input/output formats and modalities, existing methods usually design task-specific architectures and train them using face datasets collected in the particular task domain. In this work, we proposed a single model, Talk2Face, to simultaneously tackle a large number of face generation and analysis tasks, e.g. text guided face synthesis, face caption...

Relationship Analysis

Both papers belong to the Joint Understanding-Generation Architectures category, proposing unified models that perform both face analysis and synthesis tasks within a single framework. They overlap in addressing text-to-face generation, face captioning, and multi-task facial understanding through sequence-based modeling. However, UniF2ace introduces a novel Dual Discrete Diffusion (D3Diff) loss combining masked generation with score matching, employs a multi-level MoE architecture for fine-grained attribute capture, and constructs a specialized 1M-scale dataset with 17.7 attributes per caption, while Talk2Face focuses on a simpler sequence-to-sequence framework converting all tasks to natural language format with 2.3M face-text pairs across diverse existing datasets.

4. A unified deep model for joint facial expression recognition, face synthesis, and face alignment

Authors: Feifei Zhang, Tianzhu Zhang, Qi-rong Mao, Changsheng Xu, Qirong Mao | **Year/Venue**: 2020 | **URL**: View paper

Abstract

Facial expression recognition, face synthesis, and face alignment are three coherently related tasks and can be solved in a joint framework. To achieve this goal, in this paper, we propose a novel end-to-end deep learning model by exploiting the expression code, geometry code and generated data jointly for simultaneous pose-invariant facial expression recognition, face image synthesis, and face alignment. The proposed deep model enjoys several merits. First, to the best of our knowledge, this is...

Relationship Analysis

Both papers belong to the Joint Understanding-Generation Architectures category, aiming to unify face analysis and synthesis tasks within a single framework. The candidate paper focuses on jointly addressing facial expression recognition, face synthesis, and face alignment through disentangled expression and geometry codes, while the original UniF2ace paper targets fine-grained face understanding (VQA, captioning) and text-to-image generation using a unified multimodal model with discrete diffusion and MoE architecture. The key difference lies in the scope: the candidate addresses specific geometric tasks (alignment, expression recognition), whereas UniF2ace emphasizes multimodal text-image understanding and generation with fine-grained attribute control.

Contributions Analysis

Overall novelty summary. The paper proposes UniF²ace, a unified multimodal model for fine-grained face understanding and generation. It resides in the 'Joint Understanding-Generation Architectures' leaf, which contains five papers total including this one. This leaf sits within the broader 'Unified Multimodal Face Models' branch, indicating a moderately populated research direction. The taxonomy shows this is an active but not overcrowded area, with sibling works like Uniace and UniCTokens pursuing similar unification goals, though the field remains fragmented across specialized generation and analysis branches.

The taxonomy reveals neighboring research directions that contextualize this work. The sibling leaf 'General Visual Understanding-Generation Unification' contains three papers addressing broader multimodal frameworks beyond faces. Adjacent branches include 'Face Generation and Synthesis' with specialized methods for talking faces, pose synthesis, and expression modeling, plus 'Face Analysis and Recognition' focusing on understanding tasks. The taxonomy's scope notes clarify that unified models must integrate both perception and synthesis, distinguishing them from single-task approaches scattered across other branches. This positioning suggests the paper bridges traditionally separate research streams.

Among twenty-two candidates examined through semantic search, the contribution-level analysis shows varied novelty signals. The core unified model contribution examined ten candidates with none clearly refuting it, suggesting relative novelty within the limited search scope. The Dual Discrete Diffusion loss examined ten candidates and found one potentially overlapping prior work, indicating some precedent exists. The Mixture-of-Experts architecture examined only two candidates with no refutations. These statistics reflect a focused but not exhaustive literature search, leaving open questions about broader field coverage beyond top semantic matches.

Based on the limited search scope of twenty-two candidates, the work appears to occupy a moderately novel position within unified face modeling. The taxonomy structure confirms this is an emerging rather than saturated direction, though the single refutable finding for the D3Diff loss warrants closer examination of diffusion-based unification methods. The analysis covers top semantic matches and immediate taxonomy neighbors but does not claim comprehensive coverage of all relevant prior work across the fifty-paper taxonomy.

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

Contribution 1: UniF2ace: A unified multimodal model for fine-grained face understanding and generation

Description: The authors introduce UniF2ace as the first unified multimodal model that simultaneously performs both face understanding (e.g., visual question answering) and generation (e.g., text-to-image) tasks within a single framework, addressing the fragmentation in existing face research where understanding and generation are treated separately.

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. Face-makeup: Multimodal facial prompts for text-to-image generation

URL: View paper

Brief Assessment

Face-makeup[52] focuses exclusively on facial image generation using multimodal prompts (text + image), without addressing face understanding tasks like VQA or captioning. It does not claim to unify understanding and generation capabilities.

2. Facexbench: Evaluating multimodal llms on face understanding

URL: View paper

Brief Assessment

FaceXBench[53] is a benchmark for evaluating multimodal LLMs on face understanding tasks only, not a unified model for both understanding and generation. It focuses on assessment rather than proposing a unified framework.

3. Lmme3dhf: Benchmarking and evaluating multimodal 3d human face generation with lmms

URL: View paper

Brief Assessment

LMME3DHF[57] focuses on quality assessment and evaluation of AI-generated 3D human faces, not on building a unified model for both understanding and generation tasks. The candidate addresses a different problem domain (evaluation metrics) rather than unified multimodal modeling.

4. A Comprehensive Review of Unimodal and Multimodal Emotion Detection: Datasets, Approaches, and Limitations

URL: View paper

Brief Assessment

Unimodal Multimodal Emotion[59] focuses on emotion detection from facial expressions and speech for affective computing applications, not on unified face understanding and generation tasks like VQA and text-to-image synthesis that UniF2ace addresses.

5. SynAdult: Multimodal Synthetic Adult Dataset Generation via Diffusion Models and Neuromorphic Event Simulation for Critical Biometric Applications

URL: View paper

Brief Assessment

SynAdult[60] focuses on synthetic dataset generation for adult faces across multiple modalities (2D, video, neuromorphic events, 3D meshes), not on building a unified model that performs both understanding and generation tasks within a single framework.

6. A novel approach to enhancing multi-modal facial recognition: integrating convolutional neural networks, principal component analysis, and sequential neural â¦

URL: View paper

Brief Assessment

Multi-Modal Facial Recognition[58] focuses on facial recognition using CNNs, PCA, and sequential neural networks for feature extraction and classification. It does not address unified multimodal models that simultaneously perform face understanding (VQA) and generation (text-to-image) tasks within a single framework.

7. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications

URL: View paper

Brief Assessment

Multimodal Facial Expression[55] is a survey paper focused on facial expression recognition across different modalities (RGB, 3D, thermal). It does not present a unified model for face understanding and generation tasks, nor does it address the integration of these capabilities within a single framework.

8. FaceInsight: A multimodal large language model for face perception

URL: View paper

Brief Assessment

FaceInsight[54] focuses exclusively on face perception and understanding tasks (visual question answering, captioning) without any generation capabilities. The candidate does not address unified understanding-generation frameworks, making it unable to challenge the original paper's novelty claim of being the first unified model for both tasks.

9. A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges

URL: View paper

Brief Assessment

Multimodal Emotion Recognition[51] is a systematic review focused on emotion recognition from multiple modalities (facial expressions, speech, physiological signals, etc.), not on unified models for face understanding and generation tasks like VQA and text-to-image synthesis.

10. Simulated multimodal deep facial diagnosis

URL: View paper

Brief Assessment

Simulated Multimodal Diagnosis[56] focuses on medical facial diagnosis using simulated depth information for disease detection, not on unified multimodal models for general face understanding and generation tasks like VQA and text-to-image synthesis.

Contribution 2: Dual Discrete Diffusion (D3Diff) loss function with theoretical framework

Description: The authors propose D3Diff, a novel loss function that theoretically unifies score-based discrete diffusion models with masked generative models. This provides a tighter upper bound on the negative log-likelihood compared to traditional masked generative losses, enabling more precise and high-fidelity facial image generation with better alignment to fine-grained textual attributes.

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. Vector quantized diffusion model for text-to-image synthesis

URL: View paper

Brief Assessment

Vector Quantized Diffusion[67] focuses on masked generative models for discrete image tokens in VQ-VAE latent space, without proposing a unified theoretical framework combining score-based discrete diffusion with masked generative models or proving tighter bounds on negative log-likelihood as claimed in the original paper's D3Diff contribution.

2. Unified auto-encoding with masked diffusion

URL: View paper

Brief Assessment

Unified Auto-Encoding[71] focuses on combining masked auto-encoders with diffusion models for unified representation learning and generation, but does not propose a dual discrete diffusion loss that unifies score-based discrete diffusion with masked generative models as claimed in the original paper. The candidate uses standard diffusion objectives combined with masking, not a theoretical framework unifying discrete score matching with masked generation.

3. Layoutdm: Discrete diffusion model for controllable layout generation

URL: View paper

Brief Assessment

LayoutDM[65] focuses on discrete diffusion for layout generation tasks (arranging UI elements), not facial image generation. The candidate uses mask-and-replace diffusion for structured layout data, while the original proposes D3Diff to unify score-based and masked generative models specifically for high-fidelity facial image synthesis with fine-grained textual attributes.

4. Latent Wavelet Diffusion: Enabling 4K Image Synthesis for Free

URL: View paper

Brief Assessment

Latent Wavelet Diffusion[70] focuses on frequency-aware spatial supervision for ultra-high-resolution image synthesis using wavelet transforms and masking strategies. It does not address discrete diffusion models, masked generative models, or score-based discrete diffusion unification, which are the core technical contributions of D3Diff.

5. Continuously augmented discrete diffusion model for categorical generative modeling

URL: View paper

Brief Assessment

Continuously Augmented Discrete[72] focuses on augmenting discrete diffusion with continuous latent spaces to address information loss in masked tokens, rather than unifying score-based and masked generative models through a dual loss framework as proposed in the original paper.

6. Beyond masked and unmasked: Discrete diffusion models via partial masking

URL: View paper

Brief Assessment

Partial Masking[66] focuses on masked diffusion models with partial masking schemes for sub-token representations, not on unifying score-based discrete diffusion with masked generative models for facial image generation as in the original paper.

7. Di o: Distilling masked diffusion models into one-step generator

URL: View paper

Brief Assessment

Distilling Masked Diffusion[63] focuses on distilling masked diffusion models into one-step generators for efficient inference, not on unifying score-based discrete diffusion with masked generative models through a novel loss function for improved generation quality.

8. Structured denoising diffusion models in discrete state-spaces

URL: View paper

Prior Art Analysis

Structured Denoising Diffusion[64] demonstrates that prior work exists on combining score-based discrete diffusion with masked generative models. The candidate paper presents a theoretical framework proving that score-based loss (L_score) provides a tighter upper bound on negative log-likelihood compared to masked generative loss (L_2), establishing the inequality $-\log p_\theta(x_0) \leq L_1 \leq L_2$. This directly addresses the same theoretical unification that the original paper claims as novel. Both papers derive similar mathematical relationships between these two approaches and propose combined loss functions, though with different specific formulations.

Evidence

Evidence 1 - **Rationale**: Both papers propose combining masked generative objectives with auxiliary losses to improve training. While the specific formulations differ, the candidate paper's L_λ loss demonstrates the prior existence of combining these two approaches in discrete diffusion models. - **Original**: we propose the dual discrete diffusion (d3diff) loss for training posterior networks: ld3diff =- tx t=1 eq(x0)q(xt|x0) [logp θ(x0|xt)] +αlscore (pθ(x0|xt)/qt(xt|x0)) - **Candidate**: inspired by this recent work, we introduce an auxiliary denoising objective for the x0-parameterization of the reverse process, which encourages good predictions of the datax0 at each time step. we combine this with the negative variational lower bound, yielding the following alternative loss functi...

9. Cross-view masked diffusion transformers for person image synthesis

URL: View paper

Brief Assessment

Cross-view Masked Diffusion[69] focuses on pose-guided person image synthesis using masked diffusion transformers for spatial correspondence learning, not on unifying score-based discrete diffusion with masked generative models for general image generation or facial attribute control.

10. Discrete predictor-corrector diffusion models for image synthesis

URL: View paper

Brief Assessment

Discrete Predictor-Corrector[68] focuses on predictor-corrector sampling methods for discrete diffusion models, not on unifying score-based and masked generative models through a dual loss function. The candidate addresses sampling efficiency and quality through MCMC correction steps, while the original proposes a novel training objective that theoretically unifies two modeling approaches.

Contribution 3: Multi-level grouped Mixture-of-Experts architecture

Description: The authors design a hierarchical MoE architecture operating at both token-level and sequence-level, with task-specific expert groups for generation and understanding. This architecture selectively integrates semantic (CLIP) and identity (face) embeddings to address the attribute forgetting problem during representation learning, enhancing the model's ability to capture fine-grained facial attributes.

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

1. From Large Angles to Consistent Faces: Identity-Preserving Video Generation via Mixture of Facial Experts

URL: View paper

Brief Assessment

Mixture Facial Experts[62] focuses on video generation with face-specific experts (identity, semantic, detail) for identity preservation across poses, not on general multimodal understanding/generation with semantic-identity embedding integration for attribute learning as in the original paper.

2. MoDE: Mixture of Diffusion Experts for Any Occluded Face Recognition

URL: View paper

Brief Assessment

MoDE[61] focuses on occluded face recognition using diffusion experts for image reconstruction, not on integrating semantic and identity embeddings for facial attribute learning in unified multimodal models.

Appendix: Text Similarity Detection

Textual similarity detection checked 26 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. Uniace: Fine-grained Face Understanding and Generation with Unified Multimodal Models

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] UniF^{2}ace: A $\underline{Uni}$fied $\underline{F}$ine-grained $\underline{Face}$ Understanding and Generation Model View paper
- [1] Uniace: Fine-grained Face Understanding and Generation with Unified Multimodal Models View paper
- [2] UniCTokens: Boosting Personalized Understanding and Generation via Unified Concept Tokens View paper
- [3] 3D Face Modelling, Analysis and Synthesis View paper
- [4] Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset View paper
- [5] A novel generative adversarial networkâ□□based superâ□□resolution approach for face recognition View paper
- [6] Face analysis and synthesis View paper
- [7] Spacetime faces: high resolution capture for modeling and animation View paper
- [8] The face image super-resolution algorithm based on combined representation learning View paper
- [9] Super-resolving very low-resolution face images with supplementary attributes View paper
- [10] Use of a high resolution 3D optical scanner for 3D model creation, game design and facial expression recognition View paper
- [11] FG-EmoTalk: Talking head video generation with fine-grained controllable facial expressions View paper
- [12] Presentation attack detection methods for face recognition systems: A comprehensive survey View paper
- [13] Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model View paper
- [14] Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis View paper

- [15] More real than real: A study on human visual perception of synthetic faces View paper
- [16] Pisces: An Auto-regressive Foundation Model for Image Understanding and Generation View paper
- [17] Adversarial cross-spectral face completion for NIR-VIS face recognition View paper
- [18] Talk2face: A unified sequence-based framework for diverse face generation and analysis tasks View paper
- [19] Harmonizing Visual Representations for Unified Multimodal Understanding and Generation View paper
- [20] Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis View paper
- [21] Learning a high fidelity pose invariant model for high-resolution face frontalization View paper
- [22] Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures View paper
- [23] Facial analysis and synthesis scheme View paper
- [24] Fine-grained talking face generation with video reinterpretation View paper
- [25] A unified deep model for joint facial expression recognition, face synthesis, and face alignment View paper
- [26] Toward Fine-Grained Talking Face Generation View paper
- [27] Unified Autoregressive Visual Generation and Understanding with Continuous Tokens View paper
- [28] Super-Identity Convolutional Neural Network for Face Hallucination View paper
- [29] Hallucinating face by eigentransformation View paper
- [30] 3DFaceFill: An analysis-by-synthesis approach to face completion View paper
- [31] End-to-end Face-swapping via Adaptive Latent Representation Learning View paper
- [32] Facial image synthesis and super-resolution with stacked generative adversarial network View paper
- [33] Efhq: Multi-purpose extremepose-face-hq dataset View paper
- [34] Dual-agent gans for photorealistic and identity preserving profile face synthesis View paper
- [35] Analysis and synthesis of facial image sequences using physical and anatomical models View paper
- [36] 3D Facial Expressions through Analysis-by-Neural-Synthesis View paper
- [37] Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters View paper
- [38] Privacy-preserving annotation of face images through attribute-preserving face synthesis View paper
- [39] 3D Guided Fine-Grained Face Manipulation View paper
- [40] Identity aware synthesis for cross resolution face recognition View paper
- [41] Training deep face recognition systems with synthetic data View paper
- [42] Cross-spectral face completion for nir-vis heterogeneous face recognition View paper
- [43] 3D face recognition based on high-resolution 3D face modeling from frontal and profile views View paper
- [44] Facial analysis and synthesis using image-based models View paper
- [45] ANYRES: Generating High-Resolution visible-face images from Low-Resolution thermal-face images View paper
- [46] SynExpression: A Diffusion-Based Framework for Controllable Facial Expression Synthesis and Emotion Detection Using Facial Segmentation Pose Maps View paper
- [47] Face Adapter for Pre-Trained Diffusion Models with Fine-Grained ID and Attribute Control View paper
- [48] Face super resolution with a high frequency highway View paper
- [49] Fine-grained Structural Hallucination Detection for Unified Visual Comprehension and Generation in Multimodal LLM View paper
- [50] Joint Deep Learning of Facial Expression Synthesis and Recognition View paper
- [51] A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges View paper
- [52] Face-makeup: Multimodal facial prompts for text-to-image generation View paper
- [53] Facexbench: Evaluating multimodal llms on face understanding View paper
- [54] FaceInsight: A multimodal large language model for face perception View paper
- [55] Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications View paper
- [56] Simulated multimodal deep facial diagnosis View paper
- [57] Lmme3dhf: Benchmarking and evaluating multimodal 3d human face generation with lmms View paper
- [58] A novel approach to enhancing multi-modal facial recognition: integrating convolutional neural networks, principal component analysis, and sequential neural â¦ View paper
- [59] A Comprehensive Review of Unimodal and Multimodal Emotion Detection: Datasets, Approaches, and Limitations View paper
- [60] SynAdult: Multimodal Synthetic Adult Dataset Generation via Diffusion Models and Neuromorphic Event Simulation for Critical Biometric Applications View paper
- [61] MoDE: Mixture of Diffusion Experts for Any Occluded Face Recognition View paper
- [62] From Large Angles to Consistent Faces: Identity-Preserving Video Generation via Mixture of Facial Experts View paper
- [63] Di o: Distilling masked diffusion models into one-step generator View paper
- [64] Structured denoising diffusion models in discrete state-spaces View paper
- [65] Layoutdm: Discrete diffusion model for controllable layout generation View paper
- [66] Beyond masked and unmasked: Discrete diffusion models via partial masking View paper
- [67] Vector quantized diffusion model for text-to-image synthesis View paper
- [68] Discrete predictor-corrector diffusion models for image synthesis View paper
- [69] Cross-view masked diffusion transformers for person image synthesis View paper
- [70] Latent Wavelet Diffusion: Enabling 4K Image Synthesis for Free View paper
- [71] Unified auto-encoding with masked diffusion View paper
- [72] Continuously augmented discrete diffusion model for categorical generative modeling View paper