MAUGen: A Unified Diffusion Approach for Multi-Identity Facial Expression and AU Label Generation

Xiangdong Li[1], Ye Lou[1], Ao Gao[1], Wei Zhang[1,2*], Siyang Song[3*]
[1]School of Software Technology, Zhejiang University
[2]Innovation Center of Yangtze River Delta, Zhejiang University
[3]HBUG Lab, Department of Computer Science, University of Exeter
{xiangdong.li, ye_lou, gaoao.olivia, cstzhangwei}@zju.edu.cn, s.song@exeter.ac.uk

Abstract

The lack of large-scale, demographically diverse face images with precise Action Unit (AU) occurrence and intensity annotations has long been recognized as a fundamental bottleneck in developing generalizable AU recognition systems. In this paper, we propose MAUGen, a diffusion-based multi-modal framework that jointly generates a large collection of photorealistic facial expressions and anatomically consistent AU labels, including both occurrence and intensity, conditioned on a single descriptive text prompt. Our MAUGen involves two key modules: (1) a Multi-modal Representation Learning (MRL) module that captures the relationships among the paired textual description, facial identity, expression image, and AU activations within a unified latent space; and (2) a Diffusion-based Image-label Generator (DIG) that decodes the joint representation into aligned facial image-label pairs across diverse identities. Under this framework, we introduce Multi-Identity Facial Action (MIFA), a large-scale multi-modal synthetic dataset featuring comprehensive AU annotations and identity variations. Extensive experiments demonstrate that MAUGen outperforms existing methods in synthesizing photorealistic, demographically diverse facial images along with semantically aligned AU labels.

Code — https://github.com/XDLI13/MAUGen/tree/main

1 Introduction

Human Facial Action Units (AUs), defined by the Facial Action Coding System (FACS) (Ekman and Friesen 1978), encode muscle movements underlying facial expressions. They have been widely used in affective computing (Song et al. 2022), human-computer interaction (Song et al. 2025), and psychology (Donato et al. 1999) for quantifying emotional behaviors. Although reliable AU recognition models are essential for developing various real-world intelligent systems, their performance generalization, however, heavily depends on the quality, scale, and diversity of paired face–AU training data.

While recent advances in deep learning have significantly advanced facial Action Unit (AU) recognition, its progress remains hindered by the limitations of available datasets. Manual AU annotation is resource-intensive and constrained by privacy and FACS expert requirements, limiting its scalability. Consequently, widely used benchmarks such as DISFA (Mavadati et al. 2013), BP4D (Zhang et al. 2014), and CK+ (Lucey et al. 2010) contain a limited number of identities, exhibit excessive frame redundancy (e.g., near-duplicate video frames) (Hu et al. 2022), and often lack AU intensity annotations (Kollias and Zafeiriou 2019), thereby limiting their applicability for robust and fine-grained AU modeling. Large-scale in-the-wild datasets such as AffectNet (Mollahosseini, Hasani, and Mahoor 2017), EmotionNet (Fabian Benitez-Quiroz, Srinivasan, and Martinez 2016), and FABA-Instruct (Li et al. 2025) improve diversity but often suffer from weak or inconsistent labels. These limitations stem from automatic annotations using pre-trained DL models or LLMs, which are prone to errors under uncontrolled backgrounds and lack rigorous validation with pipelines shown in Figure 1. For instance, only 5% of the data in EmotionNet is used for the predictor training, raising the concern about the label reliability.

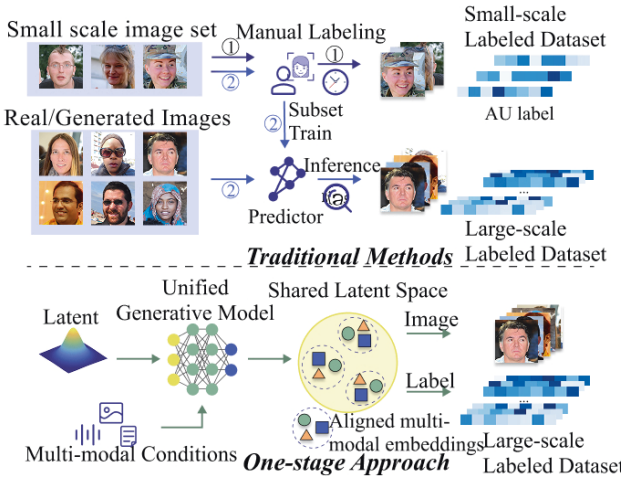

Figure 1: Comparison of two traditional AU-labeled dataset construction pipelines (indexed numerically) with our unified one-stage approach.

To address this challenge, we aim to automatically generate a large-scale, well-annotated, and demographically di-

*Corresponding authors

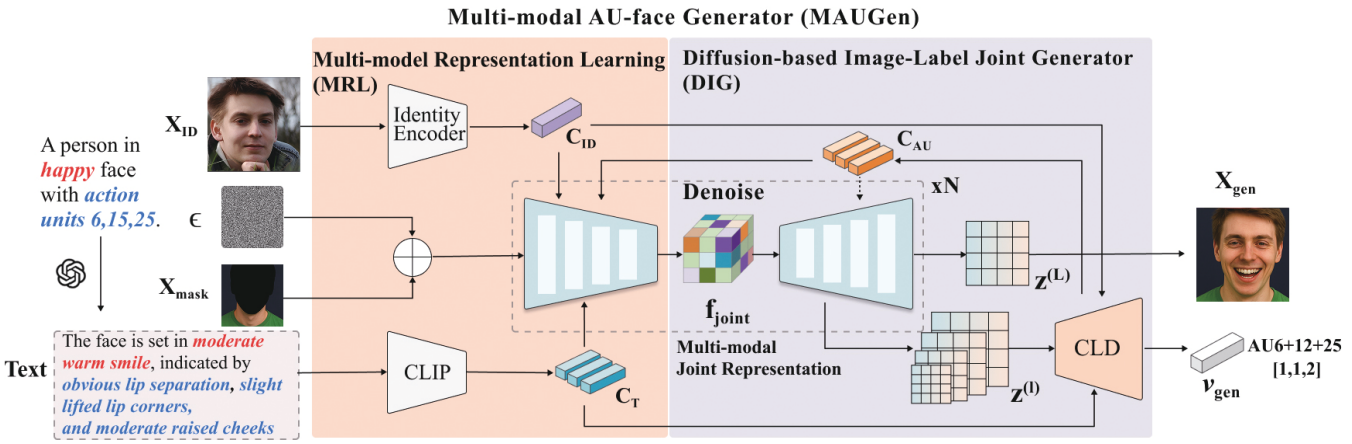

Figure 2: Overview of our MAUGen framework. The user-defined prompt is first expanded into detailed expression descriptions using an LLM. The resulting text features, together with identity exemplars, facial masks, and structure-aware AU embeddings, are encoded into a joint latent space by the MRL module (Section 3.2). The DIG module (Section 3.3) subsequently synthesizes identity-consistent facial images and corresponding AU labels via an image decoder and the Conditional Label Decoder (CLD).

verse multi-modal AU dataset. Earlier image synthesis efforts have leveraged pretrained Generative Adversarial Networks (GANs) to synthesize images with semantic masks, where predictors such as MLPs (Li et al. 2022a) or GAN inversion (Xu et al. 2023) enable the automatic annotation of large-scale unlabeled images by training on small subsets. However, such methods still rely on partially labeled data, where the trained predictors are prone to systematic errors/bias, particularly due to artifacts introduced by GAN inversion (Xia et al. 2022). Alternatively, diffusion-based image generation models such as GLIDE (Nichol et al. 2022), DALL-E 2 (Ramesh et al. 2022), Imagen (Saharia et al. 2022), and Stable Diffusion (Rombach et al. 2022a), which have unified multi-modal pipelines and adopt joint latent learning, enable one-stage co-generation of images and labels. These diffusion-based co-synthesis frameworks (Yu et al. 2023) further promote scalable and automatic dataset generation. This motivates the design of a unified one-stage pipeline for multi-modal facial behavior synthesis (see Figure 1). Nonetheless, achieving fine-grained control and high visual fidelity remains challenging, due to the ambiguity in detailed textual descriptions (Song et al. 2020) and the architectural constraints (Rosenberg et al. 2024).

In this paper, we propose MAUGen, a novel Multi-modal AU-face Generator that can jointly synthesize a large-scale dataset containing diverse facial expression images, identity-agnostic AU labels, and fine-grained textual descriptions from a user-defined prompt. It comprises two key modules: (i) the Multi-modal Representation Learning (MRL) module that constructs a unified latent space by integrating identity, text, and implicit AU structural semantics, employing a mutual conditioning strategy for cross-modal consistency; and (ii) the Diffusion-based Image-Label Joint Generator (DIG) that jointly decodes facial images and AU labels, enhanced by a triplet self-supervision mechanism and a language-guided optimization strategy. This way, our MAUGen enables accurate generation of facial expressions and their AU activations with highly detailed semantic control. The overall framework is shown in Figure 2. Our key contributions and novelties are summarized as follows:

- We propose **MAUGen**, a unified multi-modal generation framework which can jointly produce expression-rich facial images and identity-agnostic AU labels from a single textual prompt.
- We propose novel cross-modal alignment strategies in both encoding and decoding stages, enabling the generation of images with diverse expressions and identities, as well as the corresponding identity-agnostic AU labels.
- Extensive experiments show that MAUGen achieves robust multi-modal alignment. We also release the Multi-Identity Facial Action (MIFA) dataset, which contains over one million annotated images, to advance future research in AU recognition and expression analysis.

2 Related Work

Face Image Synthesis Recent advances in facial image synthesis have been primarily driven by two paradigms: generative adversarial networks (GANs) and diffusion models. GAN-based methods, such as the StyleGAN series (Karras, Laine, and Aila 2019), produce high-fidelity images with structured latent spaces conducive to editing and disentanglement. Diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022a) offer improved sample diversity and robustness. To support more precise control over facial attributes, recent efforts have introduced diverse conditioning strategies. Text-guided facial manipulation has been explored across both GAN (Patashnik et al. 2021; Xia et al. 2021) and diffusion-based (Sun et al. 2022, 2024) frameworks, with recent extensions incorporating multi-modal prompts for fine-grained expression control. Identity preservation remains a central topic, often addressed through contrastive objectives or reference-image guidance to retain subject-specific traits across expressions and styles (Tang

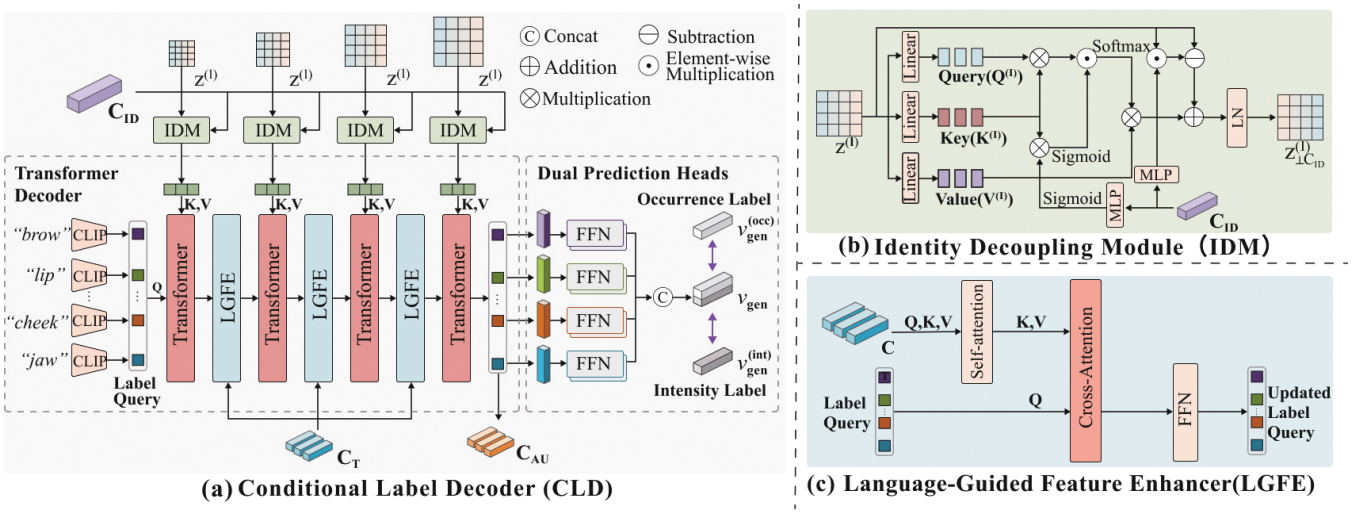

Figure 3: Structure of Conditional Label Decoder (CLD). (a) The AU queries initialized from text embeddings are refined via a Transformer Decoder with dual prediction heads and Identity Decoupling Modules (IDMs) to predict occurrence and intensity labels. (b) The IDM removes identity-related components from latent features with a modulation mask and residual filtering. (c) The Language-Guided Feature Enhancer (LGFE) injects the token-wise global semantic context into AU queries via attention.

et al. 2025; Shiohara and Yamasaki 2024). Semantic priors such as facial masks and predefined AU features have also been employed to guide localized expression synthesis (Xilin et al. 2024; Varanka et al. 2024). However, existing methods still face limitations in achieving semantically consistent expression control across diverse conditions. To address this, we propose a unified multi-modal generation framework that encodes textual semantic, identity, and AU co-activation representations into a shared latent space, enabling simultaneous synthesis of photorealistic facial images and their corresponding AU labels.

Facial Action Unit Recognition Recent advancements in AU recognition are largely driven by deep learning frameworks, including AU occurrence recognition (Yuan et al. 2024) and AU intensity recognition (Ntinou et al. 2021; Song et al. 2021). Modeling inter-AU relationships remains fundamental for capturing co-activation patterns (Luo et al. 2022; Wang et al. 2024), while generative approaches mitigate data scarcity by synthesizing AU-aware facial dynamics (Yin et al. 2024). Transformer-based architectures enhance robustness via long-range dependency modeling on RGB inputs (Liu et al. 2024). Multi-modal strategies leverage complementary cues from depth or thermal data (Zhang et al. 2024), and joint visual-semantic embeddings to enrich AU representations (Yang et al. 2021). Meanwhile, recent learning paradigms such as contrastive learning (Chang and Wang 2022), weakly supervised learning (Zhang et al. 2023), semi-supervised learning (Tang et al. 2021), and large-scale self-supervised pretraining (Ning, Salah, and Ertugrul 2024) improve the utilization of unlabeled data. Despite these advances, AU model development remains hindered by the limited scale and diversity of existing datasets, restricting generalization to complex real-world scenarios.

3 Methodology

3.1 MAUGen Pipeline

As shown in Figure 2, our MAUGen starts with applying a prompt-engineered large language model (LLM) to generate a set of diverse textual descriptions from a single textual prompt, yielding a set of facial expression descriptions detailing the activated AUs with human verification. The intensity of the AU label $v_{\text{pre}}^{(int)}$ is also incorporated in the description generation process at the training stage (details and examples are provided in *Appendix A.1*). Building on these generated descriptions, the **Multi-modal Representation Learning (MRL)** module encodes expressive semantics by integrating the expression prompt $T^{(m)}$, an identity exemplar $x_{\text{ID}}^{(n)}$ randomly sampled from an identity bank, a user-defined facial mask x_{mask} for background control, structure-aware AU cues $C_{\text{AU}}^{(m)}$ inferred during label generation (will be discussed in Section 3.3), and a noisy latent variable ϵ into a joint latent facial representation $f_{\text{joint}}^{(m,n)}$ as:

$$f_{\text{joint}}^{(m,n)} = \text{MRL}(\epsilon, x_{\text{mask}}, x_{\text{ID}}^{(n)}, T^{(m)}, C_{\text{AU}}^{(m)}) \quad (1)$$

The obtained $f_{\text{joint}}^{(m,n)}$ facilitates semantically aligned expression information, and serves as the input to the **Diffusion-based Image-label Joint Generator (DIG)** to further generate the facial expression image $x_{\text{gen}}^{(m,n)}$ and its corresponding identity-agnostic AU labels $v_{\text{gen}}^{(m)}$ as:

$$x_{\text{gen}}^{(m,n)}, v_{\text{gen}}^{(m)} = \text{DIG}(f_{\text{joint}}^{(m,n)}) \quad (2)$$

This way, the proposed MRL and DIG modules together form an end-to-end encoder–decoder backbone, enabling the joint generation of expression-rich facial images and textually consistent AU labels.

MIFA Construction Based on the proposed MAUGen, a large-scale **Multi-modal Facial AU (MIFA)** dataset is established. The MIFA contains over a million synthesized

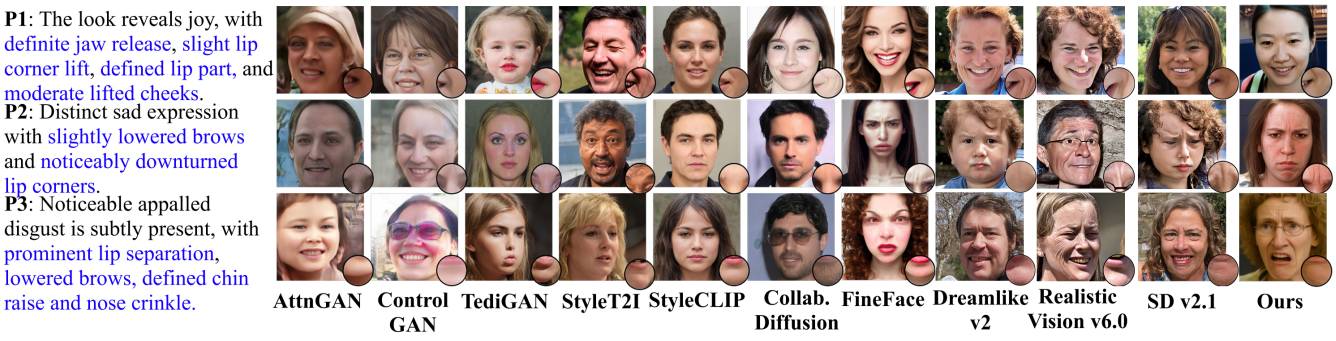

Figure 4: Qualitative comparison with prior methods. MAUGen produces photorealistic facial images with enhanced semantic alignment to textual prompts, accurately capturing fine-grained expression details. Key facial regions are highlighted.

	DISFA		BP4D	
Models	**FID**	**CLIP**	**FID**	**CLIP**
AttnGAN (Xu et al. 2018)	8.288	0.237	9.204	0.252
ControlGAN (Li et al. 2019)	7.756	0.233	9.619	0.258
TediGAN (Xia et al. 2021)	7.054	0.251	7.321	0.286
StyleCLIP (Patashnik et al. 2021)	6.899	**0.303**	6.803	<u>0.313</u>
StyleT2I (Li et al. 2022b)	7.550	0.241	<u>6.563</u>	0.268
Collab Diff (Huang et al. 2023)	<u>5.949</u>	0.238	7.992	0.285
FineFace (Varanka et al. 2024)	6.754	0.262	7.156	0.243
Dreamlike 2 (DP Team 2023)	6.919	0.231	6.955	0.277
Realistic Vision V6 (Evgeny 2024)	6.497	0.228	6.668	0.276
SD 2 (Rombach et al. 2022a)	5.741	0.229	6.659	0.275
MAUGen (Ours)	**5.516**	<u>0.295</u>	**6.517**	**0.320**

Table 1: Quantitative results on DISFA and BP4D. Models are fine-tuned per dataset. The best and second-best scores are in bold and underline, respectively.

but photorealistic and demographically diverse facial images, with each paired with 12 AU occurrence labels, 12 6-scale AU intently labels and a textual description. To guarantee the quality and reliability of our MIFA dataset, samples are filtered based on label–prompt consistency using Jaccard similarity, along with an image realism discriminator. The dataset includes 12 balanced AU labels spanning 7 primary emotions, 6 major ethnic groups, and ages ranging from children to the elderly. More details, statistics and the dataset construction pipeline are provided in *Appendix B.2–B.3*.

3.2 Multi-modal Representation Learning

As shown in Figure 2, the proposed MRL module extends the downsampling path of the *denoising UNet* (Ho, Jain, and Abbeel 2020) to incorporate multi-modal conditions, as the encoding stages of conditional diffusion models provide a stable and information-rich domain for cross-modal feature alignment. It aims to address the problem that frequently suffered by previous diffusion-based face generation models, i.e., limited semantic consistency when handling long or ambiguous textual prompts, particularly with AU details due to the absence of explicit structural priors and the lack of an effective mechanism to align AU semantics with visual representations (Shi and Fu 2025). Specifically, we propose a **Multi-modal Mutual Conditioning** strategy which introduces a series of AU co-activation representations $C_{\text{AU}}^{(m)}(t)$ $(t = 0, 1, \cdots, T)$ as auxiliary semantic priors to constrain the denoising process, where t denotes the denoising timestep. This enforces the final generated face images and AU labels to only contain plausible AU activation combinations. Here, the initial co-activation representation $C_{\text{AU}}^{(m)}(0)$ is achieved by CLIP (Radford et al. 2021) based on the description of every individual AU (examples are provided in Supplementary Material). At each denoising step t, the multi-modal joint latent facial representation $f_{\text{joint}}^{(m,n)}(t)$ is updated by a unified condition signal concatenating the textual embedding $C_{\text{T}}^{(m)}$, the identity embedding $C_{\text{ID}}^{(n)}$ extracted by pretrained identity encoder, and the AU co-activation representation $C_{\text{AU}}^{(m)}(t-1)$ encoded from the previous denoising step using the Conditional Label Decoder (CLD) (see Sec. 3.3) via cross-attention operation:

$$f_{\text{joint}}^{(m,n)}(t) \sim p\Big(f_{\text{joint}}^{(m,n)}(t-1) \mid \epsilon(x_t), C_{\text{cond}}^{(m,n)}(t-1), t\Big) \tag{3}$$

where $C_{\text{cond}}^{(m,n)}(t-1) = [C_{\text{T}}^{(m)}, C_{\text{ID}}^{(n)}, C_{\text{AU}}^{(m)}(t-1)]$. This process gradually refines the multi-modal joint latent facial representation to properly fuse face identity, AU textual description, and their co-activation constraints. The refined $f_{\text{joint}}^{(m,n)}(t)$ is, in turn, employed to update $C_{\text{AU}}^{(m)}(t-1)$ as $C_{\text{AU}}^{(m)}(t)$ within this denoising step t as:

$$C_{\text{AU}}^{(m)}(t) \sim p(C_{\text{AU}}^{(m)}(t-1) | f_{\text{joint}}^{(m,n)}(t)), \quad \text{s.t.} \quad C_{\text{AU}}^{(m)} \perp C_{\text{ID}}^{(n)} \tag{4}$$

This iterative feedback allows $C_{\text{AU}}^{(m,t)}$ to evolve from static textual priors into adaptive, context-aware cues, establishing a closed mutual conditioning loop that ensures coherent semantic alignment and facial structural consistency. To avoid circular dependencies, each denoising step computes the multi-modal joint latent facial representation $f_{\text{joint}}^{(m,n)}(t)$ using the **detached** AU co-activation representation $C_{\text{AU}}^{(m)}(t-1)$ refined from the previous denoising step. In addition, during this process, the expression cues in face identity images are mitigated through explicit structural constraints achieved by the Identity Decoupling Modules (IDMs) (see Sec. 3.3).

3.3 Diffusion-based Image-label Joint Generator

To jointly synthesize identity-varied facial images and identity-agnostic AU labels, our DIG module decodes the joint representation $f_{\text{joint}}^{(m,n)}$ via dual-branchs: one for generating facial images with diverse identities and the other for producing AU labels independent of identity cues.

Multi-identity facial expression images generation This branch uses the upsampling path of the denoising UNet to generate diverse identity-rich facial images. Specifically, we initialize the image latent z_t from the joint latent facial representation $f_{\text{joint}}^{(m,n)}$, and iteratively denoising via *Cross-Attention* conditioned on $C_{\text{AU}}, C_{\text{T}}, C_{\text{ID}}$. At each step, z_{t-1} is sampled as:

$$z_{t-1} \sim \mathcal{N}\left(\mu_\theta(z_t, [C_{\text{AU}}, C_{\text{T}}, C_{\text{ID}}]), \sigma_t^2 \mathbf{I}\right) \tag{5}$$

Here, the C_{ID} controls identity-specific facial features, while C_{AU} injects the AU co-activation patterns. μ_θ predicts the conditional mean, and σ_t^2 controls the noise level. The z_0 is decoded into the facial image x_{gen}.

Identity-agnostic AU label generation As shown in Figure $3(a)$, the label generation branch uses a Transformer-based ***Conditional Label Decoder (CLD)*** to decode AU labels from image latents $z^{(l)}$ obtained at different layers of the denoising upsampling path. Since the influence of identity-related facial characteristics may degrade accurate AU label generation, we introduce a set of ***Identity Decoupling Modules (IDMs)*** to suppress identity-related but AU-unrelated features from $z^{(l)}$ via two complementary mechanisms (shown in Figure $3(b)$) as:

- *Modulation Mask:* We define a soft attention mask as the complement to the sigmoid similarity between the projected identity and key embeddings, serving as an attention gating that downweights identity-aligned features.

$$M_{\text{ID}}^{(l)} = 1 - \text{Sigmoid}\left(\text{MLP}_{MM}^{(l)}(C_{\text{ID}}) \cdot K^{(l)\top}\right) \tag{6}$$

- *Residual Filtering:* A residual filter eliminates identity-specific cues by subtracting the latent's projection onto normalized vectors $\hat{C}_{\text{ID}}^{(l)}$, derived from C_{ID} using lightweight MLPs for dimension alignment:

$$\begin{aligned} z_{\text{res}}^{(l)} &= z^{(l)} - \langle z^{(l)}, \hat{C}_{\text{ID}}^{(l)} \rangle \cdot \hat{C}_{\text{ID}}^{(l)} \\ \hat{C}_{\text{ID}}^{(l)} &= \text{Norm}(\text{MLP}_{\text{RF}}^{(l)}(C_{\text{ID}})) \end{aligned} \tag{7}$$

Together, they decouple identity signals both at the attention level and content level, ensuring more robust AU structure learning. Although $z^{(l)}$ has deviated from the joint representation, the consistent guidance of C_{ID} throughout generation makes the suppression still effective. The purified features are computed following standard attention:

$$z_{\perp C_{\text{ID}}}^{(l)} = \text{LN}\Big(\text{Softmax}\Big(\frac{Q^{(l)}K^{(l)\top}}{\sqrt{d_k}} \odot M_{\text{ID}}^{(l)}\Big)V^{(l)} + z_{\text{res}}^{(l)}\Big) \tag{8}$$

The identity-suppressed features $z_{\perp C_{\text{ID}}}^{(l)}$ are then fused with N_q (number of AUs) learnable AU queries, initialized using CLIP encoded AU descriptions, to enable category-aware alignment. However, prompt semantics may be diluted by dominant visual signals during generation (Baltrušaitis, Ahuja, and Morency 2018). To mitigate this, we introduce a ***Language-Guided Feature Enhancer (LGFE)***, interleaved within the Transformer decoder. It applies *Self-Attention* over token-wise prompt embeddings to extract global semantic context, followed by *Cross-Attention* to inject it into AU queries, as shown in Figure $3(c)$. The refined AU queries are passed to a dual-branch prediction head comprising two independent MLPs to estimate AU occurrence $v_{\text{gen}}^{(occ)}$ and intensity $v_{\text{gen}}^{(int)}$, capturing both the activation status and variation level. Through attention-based interactions, AU queries embed co-activation patterns and structural dependencies among facial muscles. A dynamic cross-timestep Exponential Moving Average (EMA) over final-layer AU queries distills these semantics into $C_{\text{AU}}^{(m)}$, serves as a condition in subsequent representation learning.

3.4 Training Objective

Self-supervised Text–Label–Image Consistency Beyond architectural enhancements, we introduce the triplet self-supervised objectives to reinforce modality alignment: (1) *Text–Image alignment:* We adopt a local directional CLIP loss (Gal et al. 2022) to guide expression semantics. Unlike absolute similarity, which may overfit identity-preserving content, this loss emphasizes whether the semantic shift from the identity image aligns with the prompt-induced expression change; (2) *Text–Label alignment:* To align AU predictions with the expression semantics, we supervise the generated AU intensities $v_{\text{gen}}^{(int)}$ using the predefined AU vector $v_{\text{pre}}^{(int)}$ that was used to guide prompt generation:

$$\mathcal{L}_{\text{text-AU}} = \|v_{\text{gen}}^{(int)} - v_{\text{pre}}^{(int)}\|_2^2. \tag{9}$$

This encourages the label outputs to remain consistent with the expression semantics encoded in text; and (3) *Image–Label alignment:* Since prompts are constructed from predefined AU annotations in the training process, they may introduce annotation bias into the learning process. We further supervise the predicted AU occurrences and intensities using labels jointly decided by multiple open-sourced state-of-the-art AU detectors, followed by a cross-detector agreement strategy that evaluates prediction consistency across multiple auxiliary detectors, retaining only consensus labels for backpropagation. This mitigates the biases of each single AU detector. To avoid noisy supervision, AU extraction is restricted to images classified as realistic by a pretrained discriminator, ensuring structural reliability For the full formulation and details of cross-detector agreement strategy, please refer to *Appendix A.10*).

Masked Denoising Loss for Facial Regions The DDPM-style noise prediction loss (Ho, Jain, and Abbeel 2020) is applied in a spatially selective manner during latent diffusion, with a predefined facial mask x_{mask} guiding the optimization toward semantically relevant regions:

$$\mathcal{L}_{\text{mask}} = \|x_{\text{mask}} \odot \epsilon - x_{\text{mask}} \odot \epsilon_\theta(x_t, C_{\text{AU}}, C_{\text{T}}, C_{\text{ID}}, t)\|_2^2 \tag{10}$$

Total loss and Adaptive Loss Weighting Our total loss comprises a triplet loss, a masked denoising loss and an identity loss adopted from ArcFace (Deng et al. 2019). While manual weighting of multi-objective losses often requires extensive tuning and can lead to unstable/suboptimal

Metrics	Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg.
Metrics for AU Occurence Detection										
F1 ↑	FMAE (ViT-B)	51.1 / **54.6**	55.1 / **57.4**	78.6 / **72.9**	56.5 / **60.5**	44.3 / **43.2**	78.9 / **77.9**	87.8 / **86.6**	55.9 / **59.8**	63.5 / **64.1**
	FMAE (ViT-L)	47.1 / **49.5**	45.7 / **48.9**	71.4 / **74.1**	50.1 / **54.4**	49.5 / **54.6**	78.2 / **82.0**	80.6 / **88.5**	58.9 / **59.7**	60.2 / **63.9**
	FMAE (ViT-H)	44.5 / **50.6**	32.5 / **36.9**	47.6 / **75.9**	69.7 / **63.3**	46.9 / **47.3**	77.8 / **81.2**	80.4 / **91.5**	58.2 / **59.9**	57.2 / **63.3**
	GraphAU (Res50)	54.6 / **59.0**	47.1 / **41.6**	72.9 / **72.5**	54.0 / **63.5**	55.7 / **53.9**	76.7 / **83.4**	91.1 / **82.2**	53.0 / **71.8**	63.1 / **66.0**
	GraphAU (Swin-B)	52.5 / **61.9**	45.7 / **47.6**	76.1 / **74.2**	51.8 / **66.4**	46.5 / **36.7**	76.1 / **82.8**	92.9 / **86.1**	57.6 / **73.7**	62.4 / **66.2**
Metrics for AU Intensity Estimation										
ICC ↑	KJRE	.27 / **.30**	.35 / **.32**	.25 / **.22**	.51 / **.48**	.31 / **.35**	.67 / **.70**	.74 / **.76**	.25 / **.22**	.42 / **.41**
	SCC-heatmap	.73 / **.77**	.44 / **.42**	.74 / **.67**	.27 / **.30**	.51 / **.55**	.71 / **.74**	.94 / **.92**	.78 / **.82**	.64 / **.65**
	MAE-face	.73 / **.78**	.66 / **.70**	.76 / **.72**	.65 / **.68**	.60 / **.56**	.87 / **.85**	.95 / **.96**	.75 / **.69**	.75 / **.74**
MAE ↓	KJRE	1.02 / **.95**	.92 / **.97**	1.86 / **1.62**	.79 / **.85**	.87 / **.82**	.77 / **.90**	.96 / **.92**	.94 / **.98**	1.02 / **1.00**
	SCC-heatmap	.16 / **.20**	.16 / **18**	.27 / **.24**	.25 / **.28**	.13 / **.18**	.32 / **.30**	.30 / **.27**	.32 / **.35**	.24 / **.25**
	MAE-face	.11 / **.08**	.09 / **.07**	.28 / **.32**	.25 / **.24**	.13 / **.17**	.21 / **.25**	.18 / **.16**	.22 / **.28**	.18 / **.19**

Table 2: AU label consistency is assessed using F1 scores over 8 shared AUs with multiple state-of-the-art detectors. Left and bolded right values indicate results on real data from DISFA and generated samples, respectively. Intensity accuracy is measured by the Intra-class Correlation Coefficient (ICC) and Mean Absolute Error (MAE).

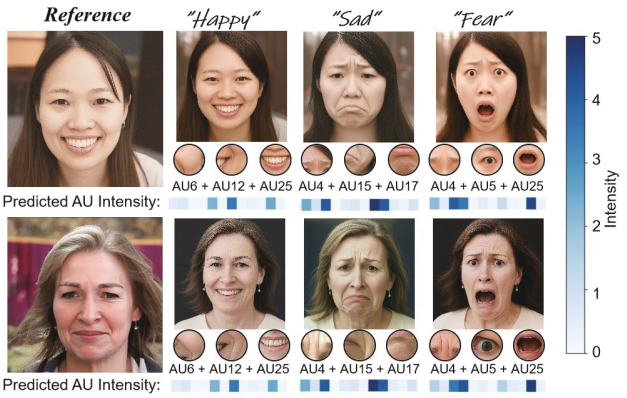

Figure 5: MAUGen-generated expressions from prompts across identities, with predicted AU intensities.

CLD	ID	MM	RF	AC	LI	LG	$\mathcal{L}_{IL}$	F1 ↑	FID ↓	CLIP ↑
								–	6.2042	0.2661
✓								92.6	6.0823	0.2425
✓	✓							88.5	6.0795	0.2237
✓	✓	✓						92.7	6.0231	0.2681
✓	✓	✓	✓					93.6	5.9120	0.2769
✓	✓	✓	✓	✓				94.2	5.8429	0.2817
✓	✓	✓	✓	✓		✓		95.4	5.8170	0.2838
✓	✓	✓	✓	✓	✓			95.7	5.6185	0.2889
✓	✓	✓	✓	✓	✓	✓		96.2	5.5772	0.2895
✓	✓	✓	✓	✓	✓	✓	✓	**96.7**	**5.5163**	**0.2949**

Table 3: Ablation Study. "MM" refers to modulation mask, "RF" to residual filtering, "MC" to mutual conditioning, "LI" to language initialization of AU queries, "LG" to LGFE, and "$\mathcal{L}_{IL}$" to the image–label alignment loss.

training due to scale mismatch across loss terms, we apply an adaptive weighting based on the relative gradient norms of each loss with respect to the final convolutional layer of the MRL module with numerical stability factor $\delta = 10^{-6}$:

$$\mathcal{L}_{\text{total}} = \sum_i \lambda_i \mathcal{L}_i\,, \quad \lambda_i = \frac{\|\nabla_\theta \mathcal{L}_i\|}{\sum_j \|\nabla_\theta \mathcal{L}_j\| + \delta} \tag{11}$$

4 Experiments

4.1 Experimental Settings

Implementation details Our MAUGen builds on Stable Diffusion (Rombach et al. 2022a,b), with the identity encoder adapted from F2D (Shiohara and Yamasaki 2024). The main training is conducted on DISFA, using both its 12 AU occurrence and intensity (ranging from 0 to 5) labels. To avoid any potential information leakage, we follow the three-fold data splitting protocol, ensuring that the training and test subsets are completely disjoint. Our training is achieved using the Adam optimizer on a single NVIDIA L20 GPU. More details on training refer to *Appendix A.2*.

Evaluation Metrics We evaluate the quality of generated face images using Fréchet Inception Distance (FID) (Heusel et al. 2017) and CLIP Score (Radford et al. 2021). The consistency of the generated AU labels and face images is measured using three variants of `FMAE` (Ning, Salah, and Ertugrul 2024) and two variants of `GraphAU` (Luo et al. 2022) AU recognition models, i.e., they predict AU occurrences from the generated face images, which are then compared with their occurrences described in the textual prompts. Mean F1 score is computed and averaged across the comparison results of five variants. This multiple variant strategy utilises the complementary strengths of different architectures, as their recognition performance varies across individual AUs. Meanwhile, `KJRE` (Zhang et al. 2019), `SCC-heatmap` (Fan, Lam, and Li 2020), and `MAE-Face` (Ma et al. 2022) are used to predict AU intensities, where intra-class correlation coefficient (ICC) and mean absolute error (MAE) are employed to compare them with intensity status described in the textual prompt.

Models	Real	Syn.	Real+Syn.
FMAE (ViT-B)	63.5	67.1	**70.9**
FMAE (ViT-L)	63.9	66.8	**70.4**
FMAE (ViT-H)	61.7	66.6	**67.9**
GraphAU (Res50)	63.1	63.5	**64.6**
GraphAU (Swin-B)	62.4	63.5	**64.1**

Table 4: F1 scores of state-of-the-art AU recognition models trained with real(R) / synthesis(S) datasets.

4.2 Qualitative Evaluation

Figure 4 presents side-by-side qualitative results across different prompts. It is clear that our MAUGen consistently generates photorealistic face images that show better alignment with the target expressions, surpassing prior methods in both visual fidelity and semantic accuracy. For example, the "expression "chin raise" in Prompt 3 is accurately rendered only by MAUGen, while other models produce ambiguous outputs with inadequate lower-face lifting. This advantage is further evidenced by complex expressions involving fine-grained AU combinations, such as AU4 (brow lowerer) combined with AU9 (nose wrinkler). As shown in Figure 5, MAUGen also excels in generating identity-agnostic expressions where the prompts are expanded from emotion keywords into detailed AU-based descriptions. The resulting outputs exhibit high inter-identity consistency in AU activation while maintaining facial identity coherence. More extensive comparisons, including failure cases and interpretative analyses, are provided in *Appendix B.5–B.6*.

4.3 Quantitative Evaluation

We compare our MAUGen with other methods that are also utilized DISFA and BP4D datasets for fine-tuning, where identity exemplars are also sampled from them for fair comparison. As shown in Table 1, MAUGen consistently achieved lower FID scores, reflecting enhanced visual realism and structural coherence, as well as competitive CLIP scores with only a slight drop on DISFA likely due to its varied background conditions (Radford et al. 2021). To assess AU label quality, we evaluate both occurrence and intensity using multiple state-of-the-art AU detectors not seen during training. As reported in Table 2, while detector-specific variations are observed, the trends largely align with the results testing on real data, further supporting the credibility of the synthesized AU labels. We also conduct controlled experiments to examine label accuracy, identity preservation, and prompt sensitivity. 5,000 identities are sampled under fixed prompts to measure identity similarity and label variance, with results detailed in *Appendix B.6–B.10*. In addition, certified FACS experts manually inspect randomly selected samples and confirm that the generated AU patterns exhibit plausible activation dynamics, with an ICC of 0.79 (see *Appendix B.3*), which is closely compared to the manual annotation accuracy on DISFA.

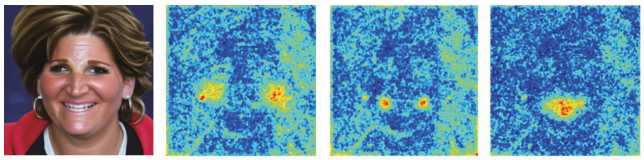

Figure 6: AU query attention maps focusing on expression-relevant regions with predicted intensities.

Models	AffectNet		Aff-Wild2	
	DISFA	MIFA	DISFA	MIFA
GraphAU (Res50)	53.21	56.01	21.45	26.72
GraphAU (Swin-B)	61.84	65.54	21.21	21.26
FMAE (Vit-B)	56.74	53.94	38.96	40.01
FMAE (Vit-H)	70.58	69.49	36.39	42.12

Table 5: Cross-dataset robustness of AU detection models trained on DISFA or MIFA.

4.4 Ablation Study

We present the main ablation experiments here, including the analysis of design components, with additional analysis of visual effect of them and adaptive loss in *Appendix B.7*.

Effect of IDM As shown in Table 3, introducing identity features without proper decoupling leads to a decrease in F1 and CLIP scores, suggesting that the identity-specific signals are inherent in label generation. Incorporating IDM significantly improves F1 from 88.5 to 93.6, alongside enhancements in CLIP and FID. Note that the F1 score is computed against the predefined AU vectors in prompt construction. We further disentangle the contributions of the modulation mask and residual filtering, both of which yield measurable gains. These results demonstrate that IDM effectively removes identity-specific signals from AU representations, resulting in more stable and identity-agnostic label generation. This effect is also visually confirmed in Figure 5, where predicted AU labels remain consistent across identities under the same prompt. We further conduct qualitative ablation studies under fixed AU prompts to examine IDM's influence, showing clearer and more consistent expressions. (visualization results in *Appendix B.7*).

Effect of Mutual Conditioning Building upon IDM, we further evaluate the effectiveness of AU-guided mutual conditioning within the CLD. As shown in Table 3, this integration leads to improvements in CLIP scores, indicating a better semantic alignment. As shown in Figure 6, final-layer attention maps reveal focused responses on expression-relevant regions, such as cheeks and lip corners. The observed co-activation across these areas suggests that the model captures structurally coherent AU patterns, validating the effectiveness of AU-guided conditioning.

Effect of Language-guided Query Optimization To evaluate the effectiveness of language-guided AU query optimization, we compare a baseline model (without query initialization and LGFE) against the full model. Quantitatively, as reported in Table 3, Language-based query initialization im-

proves the score to 95.7, and adding LGFE further raises it to 96.2. As shown in Figure 7, the t-SNE visualization of final-layer AU queries reveals that our method yields more compact and well-separated clusters, indicating enhanced discriminability of the query tokens. These improvements in evaluation scores and clustering quality confirm the effectiveness of language-guided initialization and LGFE in enhancing expression controllability.

Effect of Text-Image Alignment Mechanism As shown in Table 3, incorporating the pseudo-label loss and cross-detector agreement filtering yields consistent yet marginal improvements, confirming their ability to suppress detector-specific noise and enhance the reliability of semantic supervision. These mechanisms encourage the model to align textual cues with visual evidence more faithfully, particularly for subtle or low-intensity AUs. However, the overall gains remain moderate, likely due to the inherent domain bias of AU detectors trained on constrained datasets such as DISFA and BP4D, which limits the transferability and diversity of the supervision signals in open-domain generative settings.

Effect of the Synthesized Dataset To assess the utility of synthesized MIFA for AU recognition, we train existing AU detection models on (i) real data from DISFA, (ii) synthetic data, and (iii) a combination of both, and evaluate on real data. As shown in Table 4, models trained on the combined set consistently outperform those trained on real or synthetic data alone. These gains indicate that our synthesized samples effectively fill coverage gaps in the real dataset, particularly by introducing more diverse AU co-occurrence patterns and better-balanced distributions across both AUs and subjects (see further analysis in *Appendix B.6*).

Robustness of the MIFA Dataset The robustness of the proposed MIFA dataset was examined through a cross-dataset evaluation, which assesses its capacity to support stable performance across diverse facial expression domains. Two architectures, GraphAU with ResNet-50 and Swin-B (Luo et al. 2022), were each trained separately on the DISFA and MIFA, and both evaluated on the AffectNet and Aff-Wild2 test sets. 500 images were selected and manually annotated for each target dataset. The results listed in Table 5 show that models trained on MIFA consistently outperform those trained on DISFA on both AffectNet and Aff-Wild2, confirming MIFA's superior in supporting robust cross-domain affective analysis.

5 Conclusion

This work presents the first unified framework for multimodal facial expression data generation, capable of simultaneously synthesizing photorealistic facial images with preserved identity, identity-agnostic Action Unit (AU) labels, and descriptive textual annotations within a single generative paradigm. The proposed MAUGen framework demonstrates strong cross-modal alignment and expression diversity, leading to high-quality synthetic data that significantly boosts AU recognition performance across multiple benchmarks and model architectures. To facilitate further research, we release MIFA, a standardized and well-curated multimodal AU dataset constructed using this framework. Overall, our approach offers a scalable and extensible solution for facial expression understanding, with broad potential applications in video synthesis and fine-grained AU modeling.

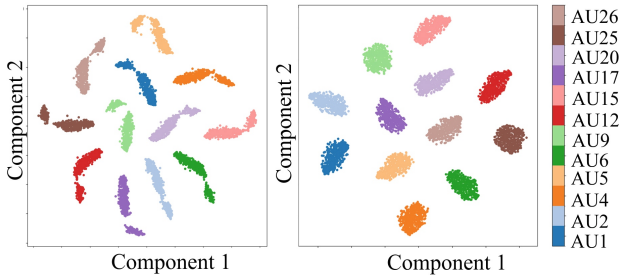

Figure 7: *t*-SNE of AU queries. *Left*: without language guidance. *Right*: with language guidance. The optimization strategy leads to more compact query clustering.

Acknowledgements

This work was supported by the Natural Science Foundation of Shaanxi Province (Grant No. 2024JC-YBMS-569) and the Key Research and Development Program of Ningbo City (Grant No. 2023Z130)

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Chang, Y.; and Wang, S. 2022. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20417–20426.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Donato, G.; Bartlett, M. S.; Hager, J. C.; Ekman, P.; and Sejnowski, T. J. 1999. Classifying facial actions. *IEEE Transactions on pattern analysis and machine intelligence*, 21(10): 974–989.

DP Team. 2023. Dreamlike Diffusion v2. https://huggingface.co/dreamlike-art/dreamlike-diffusion-2. Accessed: 2025-11-03.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Evgeny. 2024. Realistic Vision v6.0. https://huggingface.co/SG161222/Realistic_Vision_V6.0_B1_noVAE. Accessed: 2025-11-03.

Fabian Benitez-Quiroz, C.; Srinivasan, R.; and Martinez, A. M. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5562–5570.

Fan, Y.; Lam, J. C. K.; and Li, V. O. K. 2020. Facial Action Unit Intensity Estimation via Semantic Cor-

respondence Learning with Dynamic Graph Convolution. arXiv:2004.09681.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, Z.; Zhang, W.; Li, L.; Ding, Y.; Chen, W.; Deng, Z.; and Yu, X. 2022. Facial Action Units Detection Aided by Global-Local Expression Embedding. *arXiv preprint arXiv:2210.13718*.

Huang, Z.; Chan, K. C.; Jiang, Y.; and Liu, Z. 2023. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6080–6090.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kollias, D.; and Zafeiriou, S. 2019. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. 2019. Controllable text-to-image generation. *Advances in neural information processing systems*, 32.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Fidler, S.; and Torralba, A. 2022a. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21330–21340.

Li, Y.; Dao, A.; Bao, W.; Tan, Z.; Chen, T.; Liu, H.; and Kong, Y. 2025. Facial affective behavior analysis with instruction tuning. In *European Conference on Computer Vision*, 165–186. Springer.

Li, Z.; Min, M. R.; Li, K.; and Xu, C. 2022b. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18197–18207.

Liu, X.; Yuan, K.; Niu, X.; Shi, J.; Yu, Z.; Yue, H.; and Yang, J. 2024. Multi-scale promoted self-adjusting correlation learning for facial action unit detection. *IEEE Transactions on Affective Computing*.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, 94–101. IEEE.

Luo, C.; Song, S.; Xie, W.; Shen, L.; and Gunes, H. 2022. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*.

Ma, B.; An, R.; Zhang, W.; Ding, Y.; Zhao, Z.; Zhang, R.; Lv, T.; Fan, C.; and Hu, Z. 2022. Facial Action Unit Detection and Intensity Estimation from Self-supervised Representation. arXiv:2210.15878.

Mavadati, S. M.; Mahoor, M. H.; Bartlett, K.; Trinh, P.; and Cohn, J. F. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2): 151–160.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741.

Ning, M.; Salah, A. A.; and Ertugrul, I. O. 2024. Representation learning and identity adversarial training for facial behavior understanding. *arXiv preprint arXiv:2407.11243*.

Ntinou, I.; Sanchez, E.; Bulat, A.; Valstar, M.; and Tzimiropoulos, G. 2021. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Transactions on Affective Computing*, 14(1): 436–450.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2085–2094.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Rosenberg, H.; Ahmed, S.; Ramesh, G.; Fawaz, K.; and Vinayak, R. K. 2024. Limitations of Face Image Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14838–14846.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shi, L.; and Fu, Y. 2025. ExpertGen: Training-Free Expert Guidance for Controllable Text-to-Face Generation. arXiv:2505.17256.

Shiohara, K.; and Yamasaki, T. 2024. Face2Diffusion for Fast and Editable Face Personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6850–6859.

Song, S.; Jaiswal, S.; Shen, L.; and Valstar, M. 2022. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2): 829–844.

Song, S.; Sanchez, E.; Shen, L.; and Valstar, M. 2021. Self-supervised learning of dynamic representations for static images. In *2020 25th international conference on pattern recognition (icpr)*, 1619–1626. IEEE.

Song, S.; Spitale, M.; Kong, X.; Zhu, H.; Luo, C.; Palmero, C.; Barquero, G.; Escalera, S.; Valstar, M.; Daoudi, M.; et al. 2025. REACT 2025: the Third Multiple Appropriate Facial Reaction Generation Challenge. *arXiv preprint arXiv:2505.17223*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, J.; Deng, Q.; Li, Q.; Sun, M.; Ren, M.; and Sun, Z. 2022. AnyFace: Free-style text-to-face synthesis and manipulation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 18687–18696.

Sun, Z.; Song, S.; Patras, I.; and Tzimiropoulos, G. 2024. Cemiface: Center-based semi-hard synthetic face generation for face recognition. *Advances in Neural Information Processing Systems*, 37: 35612–35638.

Tang, D.; Jiang, X.; Zhang, Y.; Dai, Y.; and Lin, Y. 2025. Ipdm: identity preserving diffusion model for face sketch and photo synthesis. *Mach. Vision Appl.*, 36(2).

Tang, Y.; Zeng, W.; Zhao, D.; and Zhang, H. 2021. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12899–12908.

Varanka, T.; Khor, H.-Q.; Li, Y.; Wei, M.; Kung, H.; Sebe, N.; and Zhao, G. 2024. Towards Localized Fine-Grained Control for Facial Expression Generation. arXiv:2407.20175.

Wang, Z.; Song, S.; Luo, C.; Deng, S.; Xie, W.; and Shen, L. 2024. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1270–1280.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2256–2265.

Xia, W.; Zhang, Y.; Yang, Y.; Xue, J.-H.; Zhou, B.; and Yang, M.-H. 2022. Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3121–3138.

Xilin, H.; Cheng, L.; Xiaole, X.; Bing, L.; Siyang, S.; Muhammad, H. K.; Weicheng, X.; Linlin, S.; and Zongyuan, G. 2024. SynFER: Towards Boosting Facial Expression Recognition with Synthetic Data. arXiv:2410.09865.

Xu, A.; Vasileva, M. I.; Dave, A.; and Seshadri, A. 2023. Handsoff: Labeled dataset generation with no additional human annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7991–8000.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Yang, H.; Yin, L.; Zhou, Y.; and Gu, J. 2021. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10482–10491.

Yin, Y.; Chang, D.; Song, G.; Sang, S.; Zhi, T.; Liu, J.; Luo, L.; and Soleymani, M. 2024. Fg-net: Facial action unit detection with generalizable pyramidal features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6099–6108.

Yu, X.; Li, G.; Lou, W.; Liu, S.; Wan, X.; Chen, Y.; and Li, H. 2023. Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 592–602. Springer.

Yuan, K.; Yu, Z.; Liu, X.; Xie, W.; Yue, H.; and Yang, J. 2024. Auformer: Vision transformers are parameter-efficient facial action unit detectors. In *European Conference on Computer Vision*, 427–445. Springer.

Zhang, X.; Wang, T.; Li, X.; Yang, H.; and Yin, L. 2023. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20751–20762.

Zhang, X.; Yang, H.; Wang, T.; Li, X.; and Yin, L. 2024. Multimodal channel-mixing: Channel and spatial masked autoencoder on facial action unit detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6077–6086.

Zhang, X.; Yin, L.; Cohn, J. F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; and Girard, J. M. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10): 692–706.

Zhang, Y.; Wu, B.; Dong, W.; Li, Z.; Liu, W.; Hu, B.-G.; and Ji, Q. 2019. Joint Representation and Estimator Learning for Facial Action Unit Intensity Estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3452–3461.