LaTo: Landmark-tokenized Diffusion Transformer for Fine-grained Human Face Editing

Anonymous authors
Paper under double-blind review

Abstract

Recent multimodal models for instruction-based face editing enable semantic manipulation but still struggle with precise attribute control and identity preservation. Structural facial representations such as landmarks are effective for intermediate supervision, yet most existing methods treat them as rigid geometric constraints, which can degrade identity when conditional landmarks deviate significantly from the source (e.g., large expression or pose changes, inaccurate landmark estimates). To address these limitations, we propose LaTo, a landmark-tokenized diffusion transformer for fine-grained, identity-preserving face editing. Our key innovations include: (1) a landmark tokenizer that directly quantizes raw landmark coordinates into discrete facial tokens, obviating the need for dense pixel-wise correspondence; (2) a location-mapped positional encoding and a landmark-aware classifier-free guidance that jointly facilitate flexible yet decoupled interactions among instruction, geometry, and appearance, enabling strong identity preservation; and (3) a landmark predictor that leverages vision–language models to infer target landmarks from instructions and source images, whose structured chain-of-thought improves estimation accuracy and interactive control. To mitigate data scarcity, we curate HFL-150K, to our knowledge the largest benchmark for this task, containing over 150K real face pairs with fine-grained instructions. Extensive experiments show that LaTo outperforms state-of-the-art methods by 7.8% in identity preservation and 4.6% in semantic consistency. Code and dataset will be made publicly available upon acceptance.

1 Introduction

Generating photorealistic facial images (Lin et al., 2025) with controllable expressions, head pose, and other attributes while preserving subject identity remains a core challenge in face editing (Preechakul et al., 2022; Zhang et al., 2025). These capabilities are critical for applications such as virtual avatar creation, digital human synthesis, and identity-preserving facial modifications. Recent proprietary multimodal models (e.g., SeedEdit3 (Wang et al., 2025), Step1X-Edit (Liu et al., 2025), FLUX.1-Kontext (Labs et al., 2025)) have markedly advanced instruction-based image editing. Leveraging large-scale vision-language modeling (Li et al., 2024; Deng et al., 2025), they deliver higher-fidelity edits across diverse scenarios than prior face editing methods (Liu et al., 2022; Pernuš et al., 2023; Cheng et al., 2024). Predictably, to enable fine-grained control, users typically must provide detailed, standardized textual descriptions (e.g., "turn the subject's head 45° to the left and make the facial expression slightly happy"). However, existing models (Liu et al., 2025; Xiao et al., 2025; Labs et al., 2025) exhibit limitations in accurate instruction following and identity preservation during in-context generation. We attribute these inconsistencies to their exclusive reliance on high-level semantic encoders, which struggle to capture the structural facial cues required for precise control.

A common strategy for improving edit fidelity is to employ facial landmarks as an intermediate structural prior (Yang & Guo, 2020; Wei et al., 2025; Liang et al., 2024). Unlike text prompts, landmarks (Li et al., 2022; Sun et al., 2024) impose explicit geometric constraints via precise 2D coordinates of key facial features (eyes, nose, mouth), thereby localizing edits to the appropriate regions. However, most existing approaches are built on GANs (Goodfellow et al., 2020) or UNet-based (Ronneberger et al., 2015) diffusion models and transfer poorly to modern Diffusion Transformer (DiT) (Peebles & Xie, 2023) due to fundamental architectural differences. Recent DiT-based editors like OminiControl (Tan et al., 2024) and OmniGen (Xiao et al., 2025) adopt a general-purpose control strategy for face editing: they rasterize landmarks into 2D images, encode them via Variational Autoencoder (VAE) (Kingma & Welling, 2022) to obtain dense visual tokens, and use these as in-context guidance. Despite improving identity preservation, this strategy introduces two core limitations: (1) conditioning on rendered landmark images encourages pixel-wise copying of fixed facial shapes rather than geometric reasoning, leading to identity drift and artifacts when the conditional landmarks substantially deviate from the source in shape or position, as shown in Figure 1; and (2) because self-attention scales quadratically with sequence length (Huang et al., 2024; Avrahami et al., 2025), appending long dense visual tokens to diffusion tokens incurs prohibitive memory and compute costs, limiting practical applicability in complex scenarios.

In this paper, we present LaTo, a landmark-tokenized Diffusion Transformer for complex facial editing. Instead of relying on dense pixelwise landmark renderings, we introduce a landmark tokenizer that directly quantizes landmark coordinates

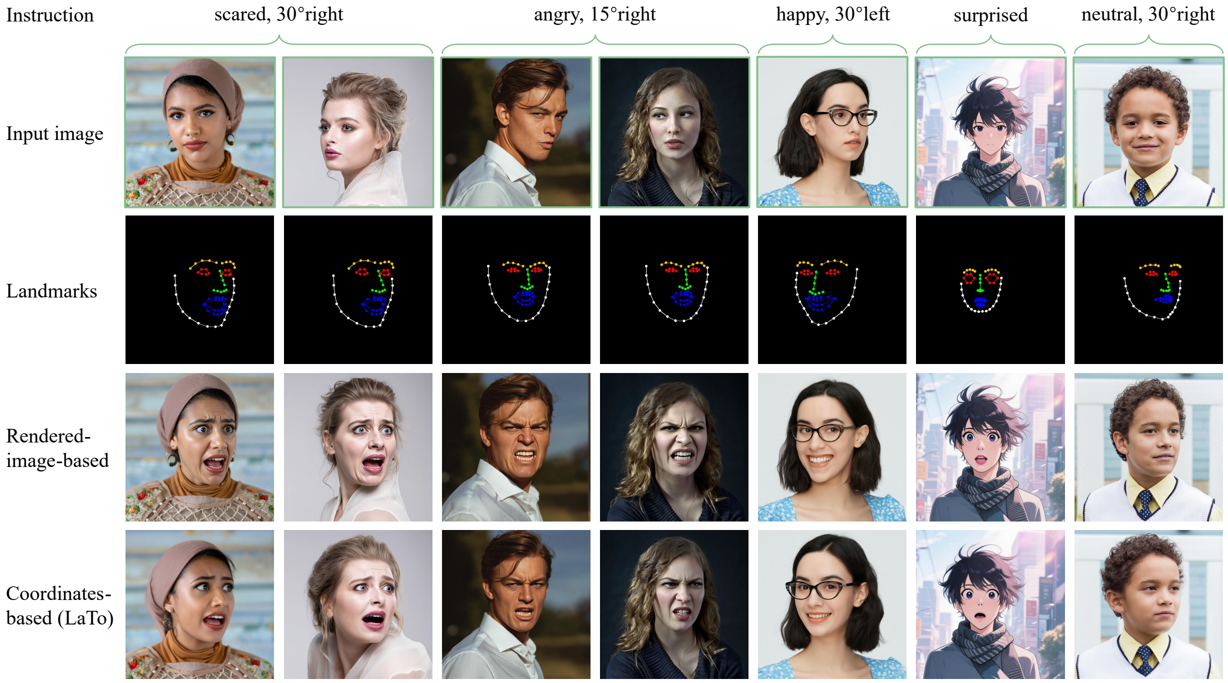


Figure 1: Landmark tokenization in LaTo preserves identity and produces natural results, whereas pixelwise alignment baselines rigidly follow the rendered landmark image and often lose identity under cross-identity landmark conditions (first four columns) or when self-identity landmarks differ substantially from the source.

into discrete facial tokens. The tokenizer adopts a VQVAE–style codebook (Van Den Oord et al., 2017), mapping coordinate inputs to embeddings with the same dimensionality as image tokens, faithfully preserving facial structure. To route sparse, spatially discontinuous landmark tokens to their target facial regions, we design a location-mapping positional encoding that anchors each token to its physical location in the latent grid, ensuring precise regional guidance in the generated image. Following Step1X-Edit, we integrate these sparse landmark tokens as contextual inputs for unified token processing within DiT blocks, enabling flexible yet decoupled interactions among geometry, appearance, and instruction while maintaining high efficiency, strong identity preservation, and semantic consistency. We further introduce landmark-aware classifier-free guidance to balance visual quality and geometric fidelity.

To address training data limitations, we develop an automated synthesis-and-curation pipeline to construct HFL-150K, a large-scale dataset of more than 150,000 face editing pairs with diverse attributes and strict identity consistency. Each pair is annotated with fine-grained editing instructions and high-precision facial landmarks, providing the rich supervision necessary to fully realize LaTo's capabilities. At inference, supplying precise landmark inputs can be impractical for end users. To alleviate this requirement, we introduce a landmark predictor that employs a vision–language model (VLM) to infer target landmarks from the source image and textual instruction using a structured chain-of-thought. We collect a set of high-quality instruction–landmark annotations with explicit change magnitudes and fine-tune a lightweight VLM, substantially improving landmark estimation accuracy and usability. Equipped with HFL-150K and the landmark predictor, LaTo achieves state-of-the-art face editing performance, particularly in semantic consistency and identity preservation. Our contributions can be summarized as follows:

- We introduce HFL-150K, a face-editing dataset comprising over 150K face pairs annotated with fine-grained editing instructions. To the best of our knowledge, it is the largest resource in this area.
- We present LaTo, the first landmark-tokenized diffusion transformer for precise face editing that integrates (i) landmark tokenization of raw coordinates, (ii) location-mapping positional encoding, and (iii) landmark-aware classifier-free guidance. This design affords flexible geometric control, strong identity preservation, and reduced computational cost compared with rendered-image conditioning.
- We develop a landmark predictor—a lightweight VLM that infers target landmarks from a source image and an editing instruction, bridging semantics and precise facial geometry for intuitive user control.
- As shown in Figure 2, leveraging HFL-150K and the Landmark Predictor, LaTo delivers precise facial attribute control under complex, fine-grained instructions and achieves state-of-the-art performance.

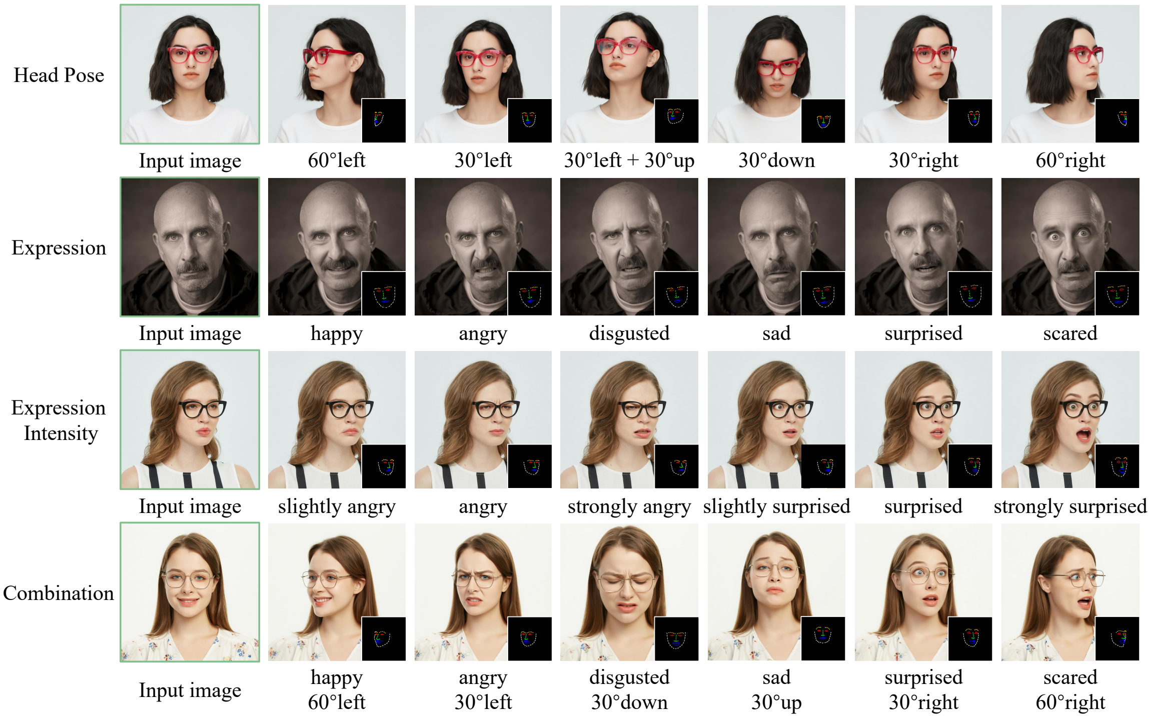


Figure 2: LaTo enables fine-grained facial expression editing, parametric head-pose editing, or their combination. The small images visualize generated landmarks via landmark predictor, enabling intuitive control signal acquisition.

2 Related Work

2.1 Instruction-based Image Editing Models

Diffusion models have become the defacto paradigm for high fidelity text to image synthesis and underpin many instruction driven editing systems. Existing approaches fall into two groups. Training-free methods manipulate the reverse process through latent inversion (Tumanyan et al., 2023; Rombach et al., 2022; Mokady et al., 2023) or attention control (Cao et al., 2023; Wang et al., 2024). These methods are efficient but often fail on complex or spatially constrained edits. Training-based methods fine tune on large scale paired image data and achieve stronger results. InstructPix2Pix (Brooks et al., 2023) pioneered synthetic supervision, while MGIE (Fu et al., 2023) and Emu Edit (Sheynin et al., 2024) incorporate VLM to improve instruction grounding. To further narrow the gap between instructions and edits, recent work couples VLM with diffusion models, including SmartEdit (Huang et al., 2024), AnyEdit (Yu et al., 2025), UltraEdit (Zhao et al., 2024), and unified frameworks such as OmniGen (Xiao et al., 2025), BAGEL (Deng et al., 2025), and ACE (Han et al., 2024). Another line fuses VLMs latents into diffusion decoders (DreamEngine (Chen et al., 2025), MetaQueries (Pan et al., 2025a), Step1X-Edit (Liu et al., 2025)). Generalist systems, for example GPT-4o (Hurst et al., 2024) and Gemini (Comanici et al., 2025), also show strong vision–language coherence. Despite these advances, precise spatial alignment and identity preservation for human face editing remain challenging because most systems rely on high level semantic signals rather than explicit geometric constraints.

2.2 Face Editing Models

Face editing seeks to modify facial attributes while preserving identity (Preechakul et al., 2022). Recent text driven approaches, including StyleCLIP (Patashnik et al., 2021) and ChatFace (Yue et al., 2023), have demonstrated strong qualitative performance. Nevertheless, they often produce entangled edits and unintended changes to identity or appearance, particularly under large instruction variations. To improve fine-grained control and anatomical consistency, subsequent work introduces structured geometric conditions such as face masks (Zhang et al., 2025), semantic layouts (Mofayezi et al., 2024), or landmark images (Li et al., 2022; Sun et al., 2024). In advanced DiT-based general editing systems (Tan et al., 2024; Pan et al., 2025a), landmarks are typically rasterized into 2D images and encoded by a visual VAE to condition the diffusion process, which strengthens geometric alignment. However, pixel-wise conditioning can encourage template copying and leads to identity drift when the target geometry differs substantially from the source. Moreover, full resolution conditionals expand the token sequence and impose high memory and computation. These limitations motivate

Table 1: Key attributes of human face editing benchmarks. HFL-150K surpasses existing face benchmarks in both scale and diversity, with unique strengths in fine-grained instruction alignment.

Benchmarks	Size	Real Image	Training	Fine-grained Instruction	Expression	Head pose
ICE-Bench (Pan et al., 2025b)	206	✓	✗	✗	✓	✗
SeqDeepFake (Shao et al., 2022)	49,920	✗	✓	✗	✓	✗
SEED (Zhu et al., 2025)	91,526	✗	✓	✗	✓	✓
HFL-150K	302,014	✓	✓	✓	✓	✓

LaTo, which directly models the relationship between landmark coordinates and target facial regions, decouples geometric structure from pixel-level appearance control.

3 Methodology

3.1 HFL-150K Dataset Construction

Large-scale editing datasets have been proven critical for developing advanced editing models. In face editing, existing datasets such as SeqDeepFake (Shao et al., 2022) and SEED (Zhu et al., 2025) suffer from two fundamental limitations: (1) they rely on coarse-grained facial attribute instructions and outdated synthesis models (Karras et al., 2019; Tsaban & Passos, 2023), resulting in unrealistic editing artifacts and limited result diversity; (2) their scale is constrained by poor-quality samples that require extensive filtering to remove invalid examples. To address the limitations of existing benchmarks, we introduce HFL-150K, a large-scale human face editing dataset comprising 150k image-edit instruction triplets (source, instruction, edited). As summarized in Table 1, HFL-150K is constructed through a hybrid approach combining real-world image curation and synthetic generation using advanced editing models.

Fine-grained attribute definition. We focus on two core facial editing tasks: expression editing (e.g., "make him smile gently") and parametric head pose editing (e.g., "rotate her head 30° left"). For expression categorization, we adopt standard emotion recognition protocols (Huang et al., 2023) to define seven canonical expressions and assign intensity levels (slightly, normally, strongly) based on visual saliency. For head pose parameters, we formulate spatial transformations using yaw and pitch angles with 30° as the base unit of motion amplitude, as shown in Figure 3 (c–d).

Synthetic data collection. As shown in Figure 3 (a), we employ advanced editing models (Step1X-Edit, GPT-4o (Hurst et al., 2024), BAGEL (Deng et al., 2025), and FLUX.1-Kontext) to generate a synthetic dataset focusing on either expression or head pose edits, aligning with their single-turn training objectives. To ensure generation quality, we implement instruction-specific filtering: (1) For expression edits, an expression validator computes semantic similarity between generated outputs and input instructions using Qwen2.5-VL (Bai et al., 2025). (2) For pose edits, a pose discriminator estimates head orientation via Euler angle regression (Yang et al., 2019) from facial landmarks and verifies alignment with target rotations θ_t, which can be described as:

$$\Delta\theta = \left\|\hat{\theta} - \theta_t\right\|_2 = \sqrt{(\hat{\theta}_p - \theta_{t,p})^2 + (\hat{\theta}_y - \theta_{t,y})^2}, \tag{1}$$

where $\hat{\theta} = (\hat{\theta}_p, \hat{\theta}_y)$ represents the estimated pitch, yaw angles. Only validated samples are retained, resulting in a **34K-sample** dataset. This synthetic dataset provides a simplified prior that enhances model interpretability.

Real-world data collection. As illustrated in Figure 3 (b), we further construct a real-world face subset from human-centric video datasets (Li et al., 2025), leveraging natural dynamics to capture intra-identity variation in expression and head pose. We apply a multi-stage filtering process comprising: (1) a quality filter using a face detector (Deng et al., 2019) to drop occlusions and enforce centering, a Laplacian of Gaussian (LoG) to remove motion blur, and Dover (Wu et al., 2023) for aesthetic and technical assessment; (2) a diversity filter combining geometric analysis (2D facial landmarks (Zhu et al., 2016)) and semantic analysis (Qwen2.5-VL for high level facial changes). Pairs with scores beyond thresholds are removed, including abnormally high values indicating copy-paste artifacts and critically low values caused by variability between different people. Finally, an image matcher with CLIP (Radford et al., 2021) performs cross frame identity verification, removing residual pairs from different people. Through this pipeline, a total of **116K** high-quality, semantically diverse image pairs are generated, reflecting natural facial variations.

For deriving fine-grained editing instructions, we leverage a facial captioner to recognize expressions and estimate intensity. However, even advanced Qwen2.5-VL-72B shows limitations in fine-grained expression recognition. To address this, we curate a high-quality dataset and fine-tune the model to improve sensitivity to more subtle magnitudes. We manually annotate 18k samples with seven predefined expression categories and intensity levels (see Appendix for guidelines). For head pose estimation, we use both horizontal and vertical optical flow angles (Karaev et al., 2024) to quantify motion amplitude relative to our base unit. A unified template is designed for instruction generation:

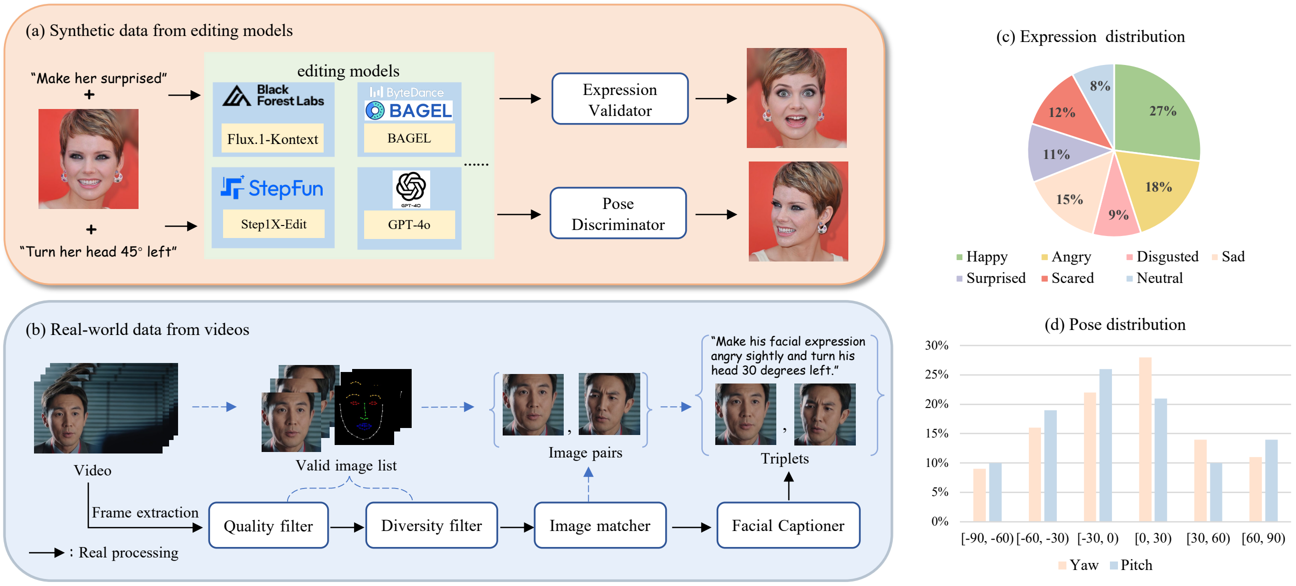


Figure 3: Data collection pipeline and statistics of HFL-150K. (a) Synthetic data generation via advanced editing models. (b) Real image pair extraction from video datasets. (c) Expression distribution across 7 categories. (d) Head pose angles aligned with 30° motion budgets.

Make his/her facial expression $\{$*expression-type*$\}$ $\{$*intensity-level*$\}$ *and turn his/her head* $\{$*angle-degree*$\}$.

3.2 LaTo

Building upon the Step1X-Edit, we propose LaTo, a fine-grained human face editing framework that effectively leverages compact facial tokens. As illustrated in Figure 4, LaTo achieves precise and user-friendly face editing through three core mechanisms: landmark tokenizer, multi-modal token fuser and landmark predictor.

3.2.1 Landmark Tokenizer

Building on the widely adopted VQVAE architecture for discrete tokenization, the landmark tokenizer combines an encoder-decoder framework with a lightweight quantizer. Given a raw landmark sequence $F = \{(X_i, Y_i)\}_{i=1}^n$ (with n 2D locations), the encoder maps it into a continuous latent space $E \in R^{n\times d}$ via residual blocks with convolutions. A quantizer then discretizes these latents through nearest-neighbor lookup in a learnable codebook $C \in R^{m\times d}$ of size m, generating compact yet expressive facial tokens in a unified geometric space. The decoder, structured to mirror the encoder, reconstructs the input sequence, ensuring spatial coherence. The complete training objective combines reconstruction loss and commitment loss:

$$\mathcal{L} = \|F - \hat{F}\|_1 + \beta\|E - \mathrm{sg}\,[C]\,\|_2^2 \tag{2}$$

where sg$[\cdot]$ denotes the stop-gradient operation, and β is the weight of the commitment loss. This loss encourages faithful reconstruction while promoting effective codebook utilization, enabling the model to learn both geometric accuracy and semantic expressiveness in facial token representations.

3.2.2 Multi-modal token fuser

We develop a token fuser to flexibly integrate landmark, image, and semantic tokens through three components: location-mapped landmark positional encoding, unified representation and landmark-aware classifier-free guidance.

Location-mapping landmark positional encoding. Step1X-Edit employs 3D Rotary Positional Encoding (RoPE) (Su et al., 2024) to encode spatial information for both image and text tokens. For each position i in image tokens, the 3D RoPE is computed as:

$$P_i = \mathrm{Concat}\left(R_T(0),\, R_H\left(\left\lfloor \frac{i}{h} \right\rfloor\right),\, R_W(i\%h)\right), \tag{3}$$

where h represents the height of the latent grid. Here, each $R(\cdot)$ implements 1D rotary embeddings applied to the text, height, and width dimensions. These embeddings are independently repeated across spatial axes to model positional rela-

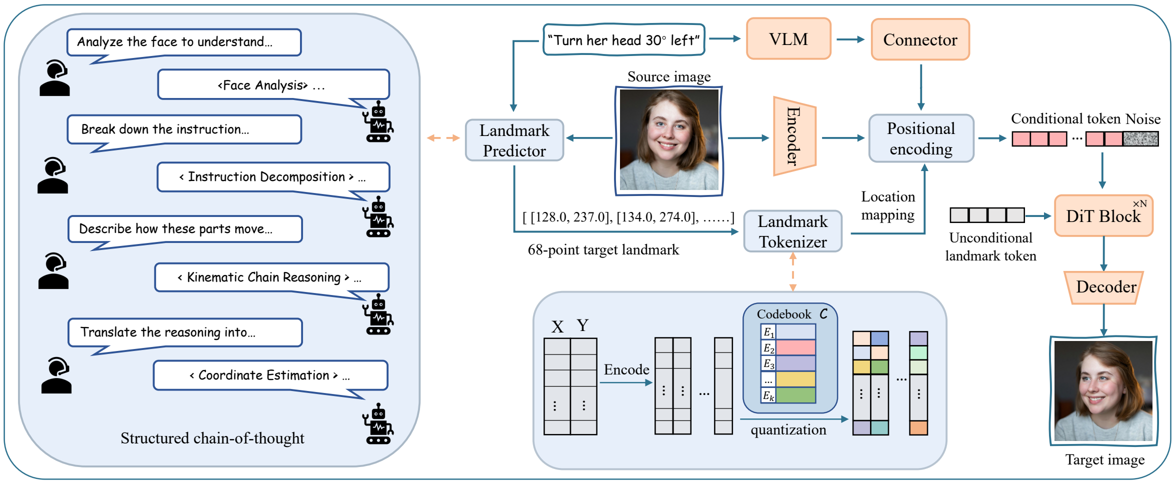


Figure 4: Overview of LaTo. The landmark predictor infers target landmarks from source image and instruction via structured chain of thought. A landmark tokenizer and visual VAE encode predicted landmarks and source image into tokens. The location-mapping positional encoding anchors each landmark token to its physical location, ensuring unified yet flexible alignment with instruction and visual tokens. The learned unconditional landmark token further guides the denoising process, keeping the edited image aligned with both the specified landmarks and instructions.

tionships between tokens. To maintain spatial consistency, we design a location mapping mechanism that links landmark tokens to their physical location in the latent grid:

$$P_i = \text{Concat}\left(R_T(0),\ R_H\left(y_i\right),\ R_W(x_i)\right), \tag{4}$$

where (x_i, y_i) denote downsampled coordinates of landmark points from the original image. This ensures each compressed representation accurately guides its corresponding area, preserving conditioning fidelity. Our experiments 4.3 demonstrate that, in the absence of this correction, the model struggles to learn spatial relationships between landmark inputs and generated tokens, leading to blurry or misaligned facial features.

Unified representation. Following Step1X-Edit, we treat all modal tokens as a unified representation. A trainable facial landmark adapter first projects facial tokens into the same latent space as noisy tokens, denoted as $z_f \in \mathbb{R}^{l_f \times d}$. The input is formulated as:

$$\mathbf{Z} = \text{Concat}(z_{\text{t}}, z_{\text{s}}, z_f, z_{\text{n}}) \in \mathbb{R}^{(l_t + l_s + l_f + l_n) \times d}, \tag{5}$$

where $z_{\text{t}} \in \mathbb{R}^{l_t \times d}$, $z_{\text{s}} \in \mathbb{R}^{l_s \times d}$, and $z_{\text{n}} \in \mathbb{R}^{l_n \times d}$ denote semantic text tokens, source visual tokens, and noisy image tokens, respectively. The multi-modal attention mechanism generates query and key via:

$$\begin{aligned} \mathbf{P} &= \text{Concat}(P_{\text{t}}, P_{\text{s}}, P_f, P_{\text{n}}) \\ \mathbf{Q} &= \text{RoPE}(W_q(\mathbf{Z}), \mathbf{P}), \quad \mathbf{K} = \text{RoPE}(W_k(\mathbf{Z}), \mathbf{P}), \end{aligned} \tag{6}$$

with $P_{\text{t}}, P_{\text{s}}, P_{\text{n}}$ derived from Equation 3 and P_f from Equation 4. This formulation enables flexible token interactions via DiT's multi-modal attention mechanism, allowing direct relationships between any token pair without rigid spatial constraints. Given the facial landmark token length $l_f = 68$ is significantly smaller than noisy image tokens ($l_n = 1024$), this approach maintains computational efficiency comparable to the baseline model.

Landmark-aware Classifier-free guidance. To balance image quality and landmark fidelity, we introduce landmark-aware classifier-free guidance (CFG). In conventional image-conditioned pipelines, the unconditional branch is obtained by feeding a zero image. By analogy, zeroing landmark coordinates for the unconditional path, especially at high CFG weights, often makes the model copy the reference face and suppress appearance variations that should covary with landmarks. We argue that zeroed landmark embeddings do not encode an unconstrained geometric state and they conflict with the physical dynamics of facial motion. This yields an undesired coupling between landmark conditions and the generated content even in the unconditional branch. Inspired by MTVCrafter (Ding et al., 2025), we train learnable unconditional tokens that replace the position-aware landmark tokens during unconditional training passes. This produces a semantically meaningful unconditional distribution and improves robustness.

3.3 Landmark Predictor

To enable intuitive interaction and improve landmark estimation, we fine-tune Qwen2.5-VL-3B into a landmark predictor (LP). Given a source image and an instruction, LP generates target landmark coordinates via a structured CoT. The CoT

Table 2: Quantitative evaluation of state-of-the-art editing methods on HFL-150K test set and face attribute editing subsets from GEdit-Bench/ICE-Bench. † Indicates models fine-tuned on HFL-150K training set.

Method	HFL-150K				GEdit&ICE-Bench(Subset)			
	SC ↑	VQ ↑	NA ↑	IP ↑	SC ↑	VQ ↑	NA ↑	IP ↑
Instruct-PixPix (Brooks et al., 2023)	0.518	0.582	0.675	0.381	0.573	0.567	0.643	0.405
AnyEdit (Yu et al., 2025)	0.612	0.641	0.702	0.446	0.669	0.654	0.676	0.479
OmniGen (Xiao et al., 2025)	0.737	0.688	0.731	0.503	0.755	0.707	0.697	0.536
Bagel (Deng et al., 2025)	0.786	0.709	0.759	0.539	0.797	0.718	0.733	0.579
Step1X-Edit (Liu et al., 2025)	0.751	0.706	0.732	0.518	0.767	0.694	0.705	0.541
Step1X-Edit† (Liu et al., 2025)	0.804	0.725	0.801	0.571	0.803	**0.732**	0.788	0.594
FLUX.1-Kontext (Labs et al., 2025)	0.712	0.720	0.779	0.556	0.749	0.693	0.751	0.576
FLUX.1-Kontext† (Labs et al., 2025)	0.786	0.737	**0.816**	0.593	0.801	0.713	0.771	0.609
LaTo (ours)	**0.832**	**0.749**	0.805	**0.634**	**0.829**	0.724	**0.793**	**0.651**

supervision comprises four stages: (1) initial state analysis, characterizing the starting pose, expression, and landmark alignment; (2) instruction decomposition, breaking the instruction into primary anatomical motions; (3) kinematic-chain reasoning, separating rigid motions (head rotation, translation) from non-rigid deformations (muscle-driven expression changes); and (4) coordinate estimation, mapping these components to numerical displacements on a normalized 512×512 canvas and producing a canonical, machine-parsable list of (X, Y) pairs. We generated CoT traces for 23,145 triplets sampled from HFL-150K using a rule-guided pipeline that ingests the source image, instruction and target landmarks, and after manual verification retained 19,398 high-quality examples for fine-tuning. During training, the visual and textual inputs are encoded and fused in the multimodal transformer and the model is optimized with next-token supervision to produce the structured CoT token sequence. Coordinates are normalized and encoded with a compact tokenization scheme and a fixed output grammar to improve numeric fidelity. At inference, smoothing and geometric sanity checks are applied to preserve identity-consistent rigid distances. This design delivers interpretable, numerically precise landmark predictions that connect robust instruction understanding to explicit geometric control for downstream face editing. Detailed CoT procedures are provided in the Appendix.

4 Experiments

4.1 Experimental Setup

Implement details. The landmark tokenizer is trained from scratch on HFL-150K with an 8,192-entry codebook and 3,072-dimensional codes matched to the Step1X-Edit hidden size. Training uses 8 NVIDIA A100 GPUs with batch size 128 per GPU for 100k iterations, and unused codes are reset every 50 steps to prevent saturation. For LaTo, we fine-tune the base model with LoRA (rank 64) and add an unfrozen linear landmark adapter. The model is trained on 16 NVIDIA A100 GPUs with total batch size 32, learning rate 1×10^{-4}. We replace conditional landmark tokens with unconditional ones with probability 0.1 to encourage diversity, and train for a total of 40k iterations.

Dataset. We split HFL-150K into 150,000 training samples and 1,007 test samples. For the test set, stratified sampling ensures diverse coverage of expression types, head poses, and their combinations. We additionally perform manual validation to ensure all samples include high-quality instructions and image–landmark pairs. We also curate an auxiliary test set by selecting face-attribute editing samples from GEdit-Bench (Liu et al., 2025) and ICE-Bench (Pan et al., 2025b), yielding 103 samples for further evaluation. Detailed curation protocols are provided in the Appendix.

Metrics. We propose a four-criteria framework to evaluate both identity preservation and accuracy of requested modifications, comprising Semantic Consistency (SC), Visual Quality (VQ), Natural Appearance (NA), and Identity Preservation (IP). SC, VQ, and NA are scored by Qwen2.5-VL-72B using a normalized $[0, 1]$ scale via visual reasoning. For IP, we first use ArcFace similarity s_{arc} (Deng et al., 2019), but some methods inflate it by copying or barely altering the source. We therefore define a rectified score that penalizes such cases: Qwen2.5-VL parses the source face and predicts expected edit amplitude $\varphi_{ins}$from the instruction, while the actual amplitude φ_{real} is derived from SSIM between source and edit. The rectified IP score calculated as:

$$s_{rip} = \max(0, s_{arc} - (\frac{\varphi_{ins} - \varphi_{real}}{\varphi_{ins} + \epsilon})^2), \tag{7}$$

we provide the pseudocode in the Appendix and confirm that this metric achieves better alignment with human preference through a user study.

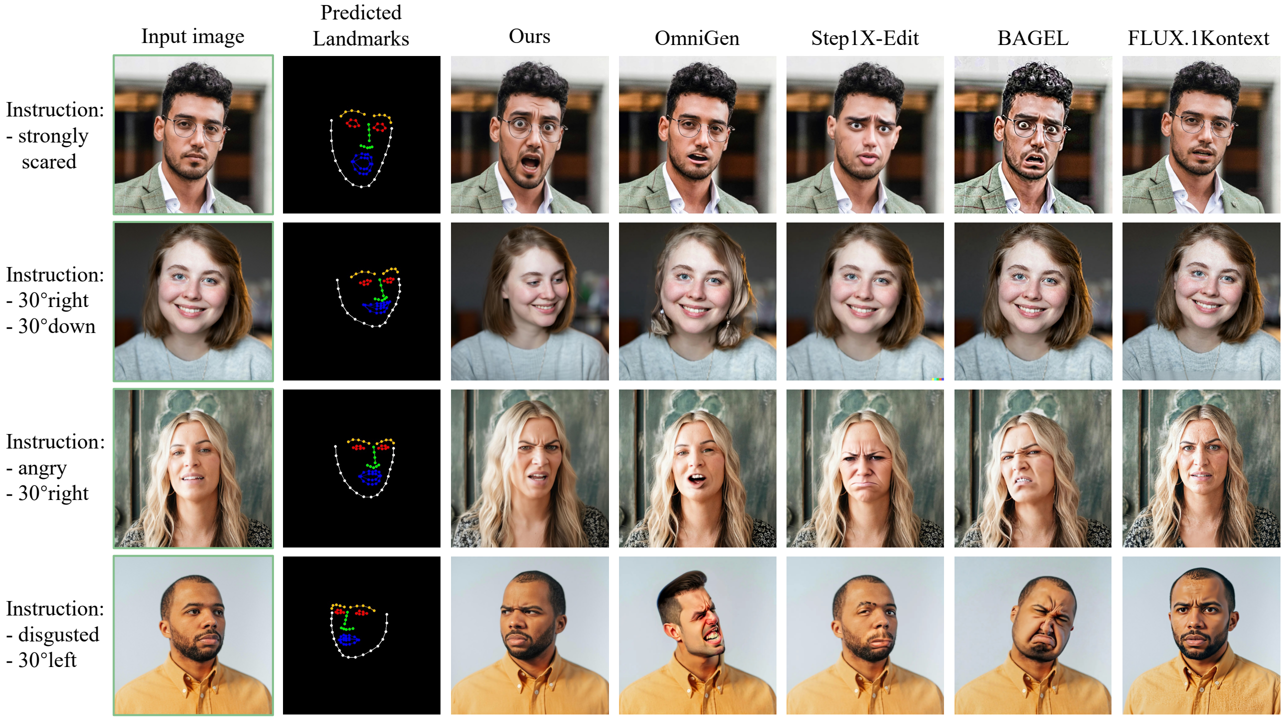


Figure 5: Qualitative comparison with state-of-the-art image editing methods.

Table 3: Ablation study on landmark condition types and landmark positional encoding. We calculate the L1 distance between the edited image and provided landmark as the Landmark Error to evaluate the landmark alignment precision.

Landmark type	Encoder	Additional Setting	SC ↑	VQ ↑	NA ↑	IP ↑	Latency(s) ↓	Landmark Error ↓
/	/	/	0.804	0.725	0.801	0.571	**49.6**	-
rendered image	visual VAE	/	0.821	0.712	0.744	0.584	83.6	**1.76**
		compression	0.816	0.704	0.709	0.569	61.3	3.07
coordinates	landmark tokenizer	(1) w/o PE	0.654	0.611	0.630	0.512	50.7	58.9
		(2) RoPE	0.778	**0.751**	0.786	0.621	52.1	25.4
		(3) learnable RoPE	0.803	0.737	0.791	0.617	52.9	9.63
		(4) ours	**0.832**	0.749	**0.805**	**0.634**	52.1	2.34

4.2 Qualitative and Quantitative Analysis

We benchmark LaTo against state-of-the-art methods on two datasets: HFL-150K (fine-grained edit instructions) and the face-attribute splits of GEdit-Bench/ICE-Bench (global descriptions). For fairness, we fine-tune Step1X-Edit and Kontext on HFL-150K using their official implementations. Table 2 reports statistically significant gains across all metrics. We observe 5.3% and 7.4% relative improvements in SC on HFL-150K for the two approaches, suggesting that our dataset better reflects real-world diversity for this task. On HFL-150K, LaTo surpasses the second-best method, Bagel, by 4.6% SC, and LaTo-IP exceeds FLUX.1-Kontext by 7.8%. On GEdit-Bench/ICE-Bench, LaTo outperforms Bagel by 7.2%, demonstrating stronger identity preservation. Despite sharing the same training data and base models as Step1X-Edit†, LaTo achieves a 2.9% average absolute improvement across all metrics, validating the effectiveness of our landmark tokenization design. Qualitative results (Figure 5) show superior pose accuracy and identity consistency, while maintaining photorealism under large expression changes where most baselines introduce cartoon-like or synthetic artifacts.

4.3 Ablation study

Facial condition formulations. We compare against two standard landmark conditioning schemes: (1) rendering landmarks as 2D images and extracting facial tokens with a shared VAE, and (2) downsampling landmark images with position shifting (inspired by OmininControl2 (Tan et al., 2025)) to improve efficiency. As shown in Table 3, relative to the fine-tuned baseline, these image-based variants are limited: NA decreases by 5.7%, IP increases only by 1.3%, and compression

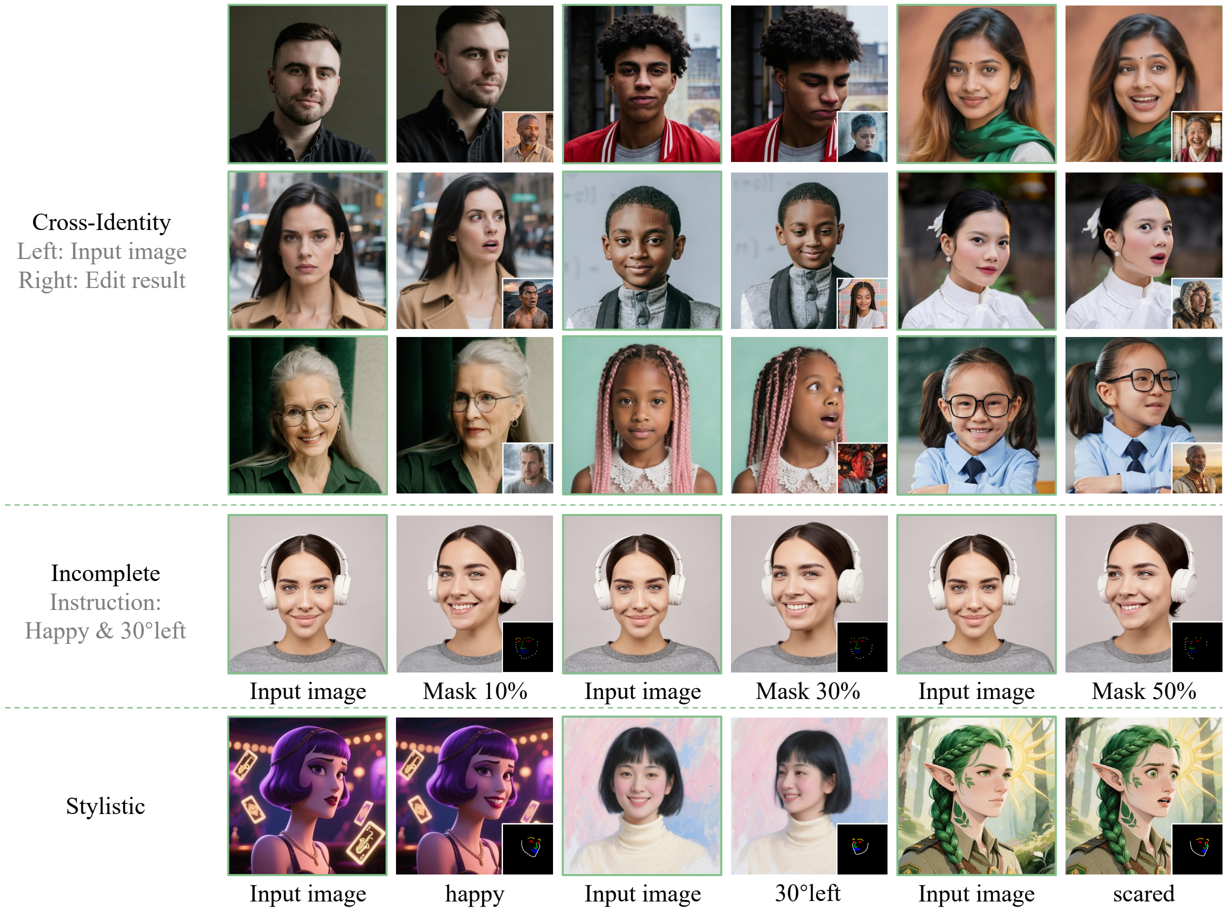


Figure 6: Qualitative results on challenging inputs, including cross-identity landmarks, incomplete landmarks, and stylistic inputs. For cross-identity cases, the corresponding driving images are shown in the bottom-right corner.

Table 4: Comparison of landmark predictor performance using unified prompting conditions.

Methods	Accuracy
Qwen2.5-VL-3B	0.477
Qwen2.5-VL-72B	0.597
Gemini 2.5 Pro	0.613
Ours	**0.730**

Table 5: Performance analysis across different CFG scales for landmark condition.

CFG-scale	SC ↑	VQ ↑	NA ↑	IP ↑	Landmark Error ↓
1	0.803	0.733	**0.816**	0.619	9.63
4	**0.832**	**0.749**	0.805	**0.634**	2.34
7	0.821	0.698	0.751	0.616	1.94
10	0.807	0.656	0.679	0.588	**1.82**

further widens the gap despite a 26% speedup. In contrast, our landmark tokenization models spatial relations between coordinates and facial attributes and decouples control strength, achieving 6.1% higher NA, 1.1% higher SC, and 5.0% higher IP than rendered-image conditioning. It also attains a 37% speedup, matching the baseline's computational efficiency.

Landmark positional encoding effectiveness. We investigate various positional encoding (PE) strategies for landmark tokens, including original relative encoding (RoPE), learnable RoPE, no PE, and our location-mapping RoPE. The results in Table 3 demonstrate that: (1) No PE leads to unstable training and unnatural results due to the absence of spatial awareness; (2) Original RoPE fails to effectively capture spatial relationships between landmarks and target images, achieving a suboptimal landmark error of 25.4; (3) Learnable RoPE improves both instruction adherence and landmark conditioning but remains inferior to our approach; (4) our method provides the model with a strong geometric prior, enabling rapid geometry information extraction and achieving superior performance in both landmark conditioning and identity fidelity.

Landmark predictor accuracy evaluation. To evaluate the effectiveness of landmark predictor, we conducted a manual accuracy assessment against Gemini 2.5 Pro, Qwen2.5-VL-72B, and Qwen2.5-VL-3B. Specifically, we selected 50 human faces from the HFL-150K test set and randomly generated instructions for each image by altering expression, head pose angle, or their combinations (6 variations per image), resulting in 300 samples. The predicted landmarks were rendered on

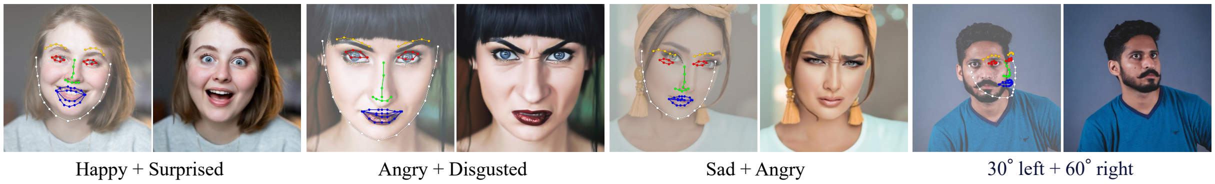


Figure 7: Visualization of the performance of landmark predictor under ambiguous instructions. Predicted landmarks are overlaid on the source images.

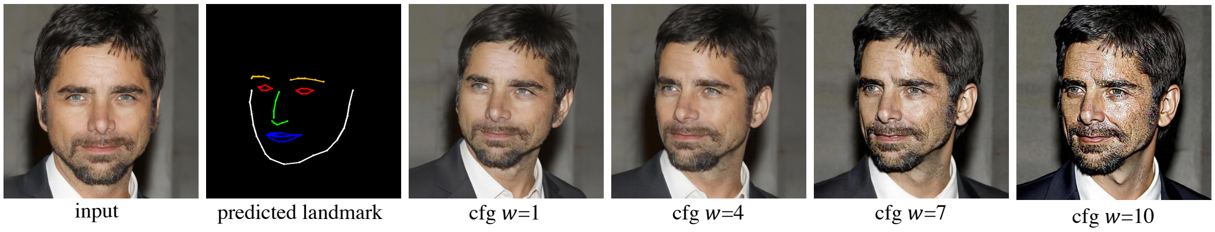


Figure 8: Visualization of the landmark-aware CFG scale w.

source images, and participants were asked to evaluate whether the predicted landmarks aligned with the given instructions, assigning a score of 0 (incorrect) or 1 (correct). We recruited 10 human evaluators and collected their results, which are presented in Table 4. Our fine-tuned model achieves the highest average accuracy among the compared models, outperforming the baseline by 25.3%, demonstrating its superiority in landmark analysis. To further evaluate LP under realistic noisy instructions, we test it on composite expressions (e.g., happy + surprised) and conflicting pose instructions (e.g., turn left and turn right), as shown in Figure 7. LP remains robust to instruction-level ambiguity. For composite expressions, it produces plausible intermediate states in the landmark space. For conflicting poses, the CoT module infers a coherent head orientation and often defaults to a near-frontal view when cues conflict.

Landmark-aware CFG scaling analysis. Figure 8 and Table 5 illustrate the impact of our CFG scale. Raising the CFG scale improves landmark alignment, yet may generate more artifacts and compromise video quality. We adopt a CFG scale of 4 as the optimal baseline configuration.

The sensitivity to challenging inputs. Although LaTo is primarily self-landmark-driven via the proposed LP, we also evaluate its behavior in three challenging scenarios: (1) incomplete landmarks, (2) cross-identity landmarks, and (3) stylistic inputs. As shown in Figure 6, LaTo largely preserves identity under cross-identity conditioning and moderate landmark dropout, and maintains stable editing quality on non-photorealistic inputs. We attribute this to three key components. First, the landmark tokenizer disentangles explicit shape-related structure from pixel-level appearance, which stabilizes geometric reasoning and facilitates the capture of generalized expression dynamics and identity-agnostic pose variations . Second, the token fuser encourages the model to interpret driving facial geometry, editing instructions, and visual identity features in a coordinated manner, which enables driving-geometry control in an implicit, feature-level manner, rather than relying on explicit geometric following. Third, the learned unconditional landmark tokens give each landmark position a flexible way to adjust the coupling between geometry and identity, so that even when landmark absence becomes severe (about 50%), LaTo, despite noticeable degradation in identity preservation, still maintains basic visual quality and key facial traits.

5 Conclusion

We presented LaTo, a landmark-tokenized diffusion transformer for fine-grained, identity-preserving face editing. LaTo quantizes landmark coordinates into discrete facial tokens and aligns them with image tokens via a location mapping positional encoding, which decouples geometry from appearance and enables precise control with strong identity preservation and high efficiency. A vision–language landmark predictor with structured reasoning infers target landmarks from instructions and source images, improving robustness and interactive controllability. This design removes the need for dense pixel-wise correspondence and mitigates identity drift under large pose or expression changes. To support research at scale, we curate HFL-150K, a large-scale benchmark of face pairs with fine-grained instructions, spanning real-world imagery and outputs from advanced models. Extensive experiments demonstrate that LaTo delivers state-of-the-art photorealism, semantic consistency, and computational efficiency, establishing a strong foundation for controllable, human-centric editing.

References

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7877–7888, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22560–22570, 2023.

Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. *arXiv preprint arXiv:2502.20172*, 2025.

Yuhao Cheng, Zhuo Chen, Xingyu Ren, Wenhan Zhu, Zhengqin Xu, Di Xu, Changpeng Yang, and Yichao Yan. 3d-aware face editing via warping-guided latent direction learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 916–926, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yanbo Ding, Xirui Hu, Zhizhi Guo, Chi Zhang, and Yali Wang. Mtvcrafter: 4d motion tokenization for open-world human image animation, 2025. URL `https://arxiv.org/abs/2505.10238`.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language modelbrooks2023instructpix2pixas. *arXiv preprint arXiv:2309.17102*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2024.

Zi-Yu Huang, Chia-Chin Chiang, Jian-Hao Chen, Yi-Chian Chen, Hsin-Lung Chung, Yu-Ping Cai, and Hsiu-Chuan Hsu. A study on computer vision for facial emotion recognition. *Scientific reports*, 13(1):8425, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos, 2024.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4176–4185, 2022.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7752–7762, 2025.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. URL `https://arxiv.org/abs/2405.08748`.

Jiayi Liang, Haotian Liu, Hongteng Xu, and Dixin Luo. Generalizable face landmarking guided by conditional face warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2425–2435, 2024.

Li Lin, Santosh Santosh, Mingyang Wu, Xin Wang, and Shu Hu. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3503–3515, 2025.

Kanglin Liu, Gaofeng Cao, Fei Zhou, Bozhi Liu, Jiang Duan, and Guoping Qiu. Towards disentangling latent space for unsupervised semantic face editing. *IEEE Transactions on Image Processing*, 31:1475–1489, 2022.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Mohammadreza Mofayezi, Reza Alipour, Mohammad Ali Kakavand, and Ehsaneddin Asgari. M^3 face: A unified multi-modal multilingual framework for human face generation and editing. *arXiv preprint arXiv:2402.02369*, 2024.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025a.

Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Ice-bench: A unified and comprehensive benchmark for image creating and editing, 2025b. URL `https://arxiv.org/abs/2503.14482`.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2085–2094, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. Maskfacegan: High-resolution face editing with masked gan latent code optimization. *IEEE Transactions on Image Processing*, 32:5893–5908, 2023.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation. In *European Conference on Computer Vision*, pp. 712–728. Springer, 2022.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. Lafs: Landmark-based facial self-supervised learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1639–1649, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers, 2025. URL `https://arxiv.org/abs/2503.08280`.

Linoy Tsaban and Apolinário Passos. Ledits: Real image editing with ddpm inversion and semantic guidance. *arXiv preprint arXiv:2307.00522*, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing, 2025. URL `https://arxiv.org/abs/2506.05083`.

Mengting Wei, Tuomas Varanka, Yante Li, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Towards consistent and controllable image synthesis for face editing. *arXiv preprint arXiv:2502.02465*, 2025.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20144–20154, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1087–1096, 2019.

Yang Yang and Xiaojie Guo. Generative landmark guided face inpainting. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 14–26. Springer, 2020.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

Dongxu Yue, Qin Guo, Munan Ning, Jiaxi Cui, Yuesheng Zhu, and Li Yuan. Chatface: Chat-guided real face editing via diffusion latent space manipulation. *arXiv preprint arXiv:2305.14742*, 2023.

Xin Zhang, Siting Huang, Xiangyang Luo, Yifan Xie, Weijiang Yu, Heng Chang, Fei Ma, and Fei Yu. Museface: Text-driven face editing via diffusion-based mask generation approach. *arXiv preprint arXiv:2503.23888*, 2025.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3409–3417, 2016.

Yule Zhu, Ping Liu, Zhedong Zheng, and Wei Liu. Seed: A benchmark dataset for sequential facial attribute editing with diffusion models. *arXiv preprint arXiv:2506.00562*, 2025.

Appendix

A The Use of Large Language Models

After completing the manuscript, we used large language models to check grammar and refine the writing.

B Dataset Illustration

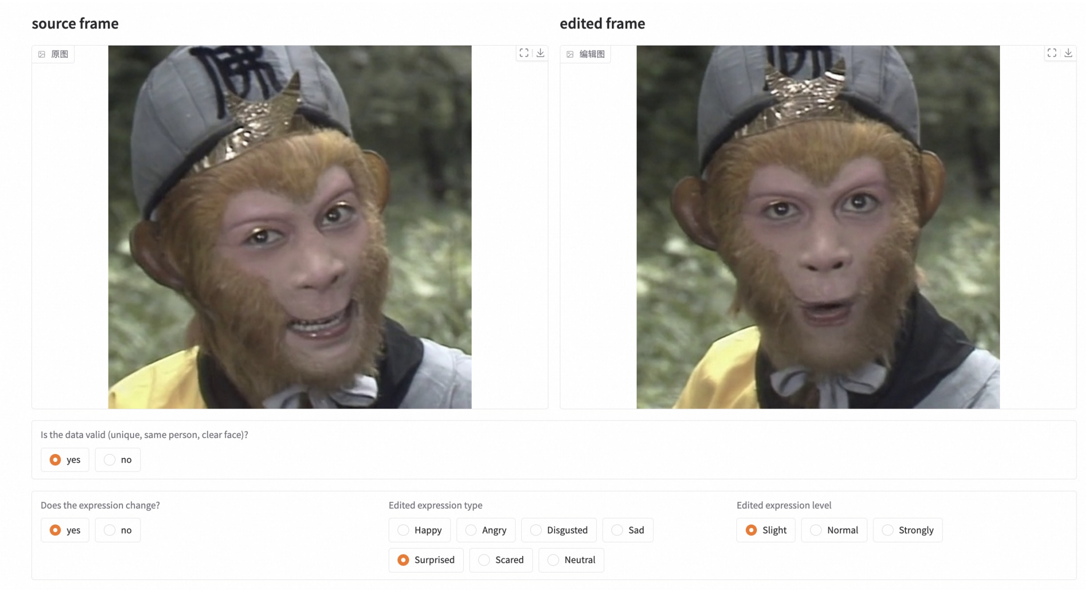
B.1 More Details of filtering pipeline

For real image pairs, the quality filter retains samples that contain a single face whose centroid lies within the central 20% of the image, with facial landmarks covering at least 7% of the image area, and with motion blur below a Laplacian of Gaussian (LoG) threshold of 50. Dover is applied with a threshold of 0.5 for aesthetic and technical assessment. The diversity filter then enforces both geometric and semantic coverage. For geometric diversity, we compute a composite change score combining local (eyes/eyebrows/mouth; indices 36–47 and 48–67 in the 68-point scheme) and global landmark displacements, normalized by inter-ocular distance, with a threshold of ≥ 23:

$$\text{change score} = 0.7 \times \text{inner diff} + 0.3 \times \text{overall diff}, \tag{8}$$

where inner diff is the mean absolute difference of the inner facial landmarks (indices 36–47 and 48–67), and overall diff is the mean absolute difference across all 68 landmarks. For semantic diversity, we employ Qwen2.5-VL-32B to calculate a [0,1] normalized difference score between high-level visual features, retaining only pairs exceeding a threshold of 0.4. Finally, CLIP (Radford et al., 2021) is used for cross-frame identity verification, with a high similarity threshold of 0.9. For synthetic data, we estimate pitch and yaw from 68-point facial landmarks and retain only samples whose angular deviation from the instructed pose is within 10°.

B.2 Expression instruction for real images



Is the data valid (unique, same person, clear face)?

● yes ○ no

Does the expression change?

● yes ○ no

Edited expression type

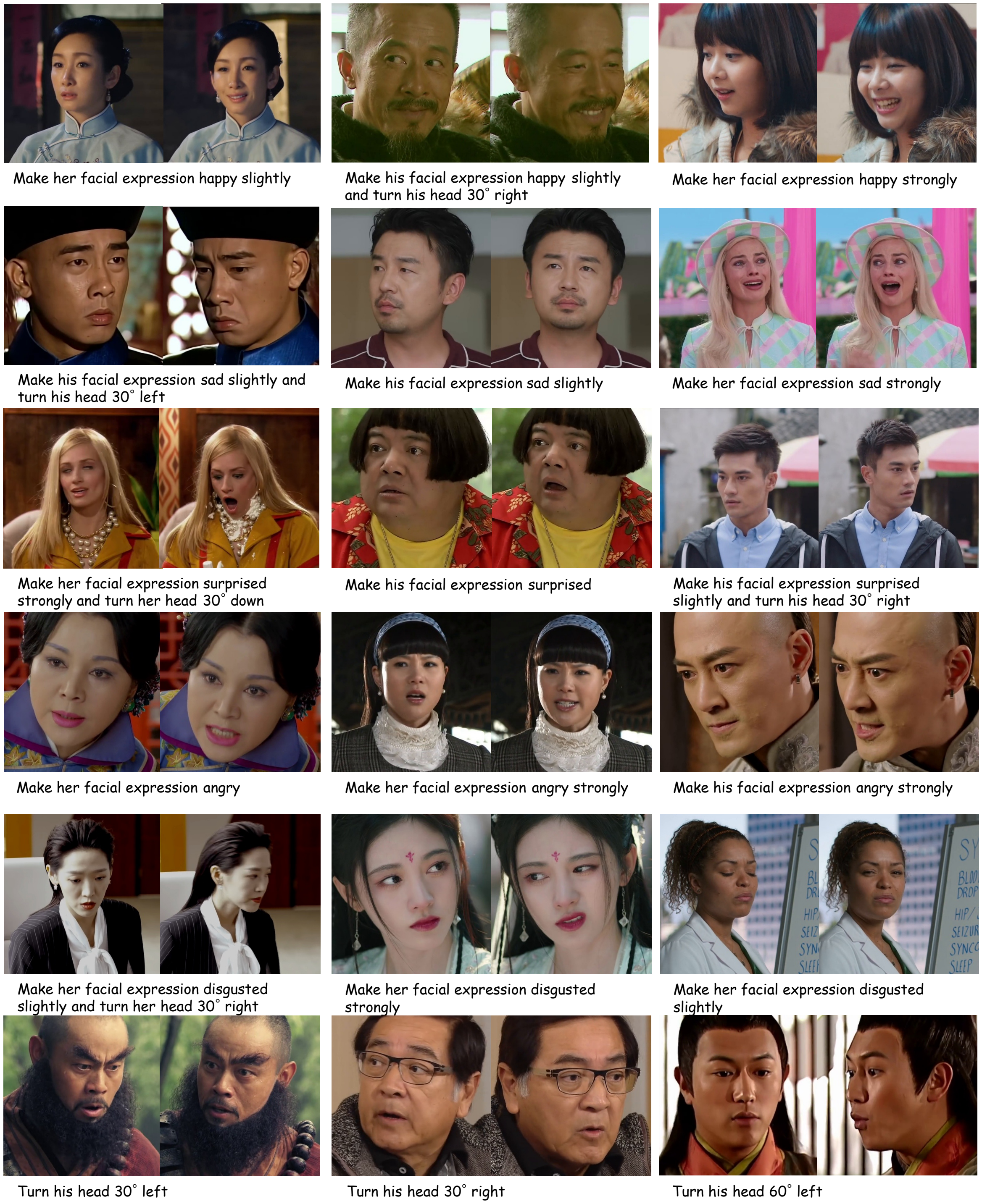
○ Happy ○ Angry ○ Disgusted ○ Sad ● Surprised ○ Scared ○ Neutral

Edited expression level

● Slight ○ Normal ○ Strongly

Figure 9: User interface for expression annotation.

We observe that even Qwen2.5-VL-72B struggles to capture expression changes with sufficient granularity. To address this, we curate an 18K expression caption dataset with intensity labels via a crowdsourcing interface (Figure 9). Five annotators independently score each sample, and we retain only items with agreement from at least three raters. This corpus is used to fine-tune Qwen2.5-VL-7B, with 17K samples for training and 1K for evaluation. Evaluation adopts a hierarchical metric. Expression recognition receives 1 point only when the predicted category matches the ground truth. Intensity estimation receives 0.5 points only when the expression is correct and the predicted intensity matches the annotation. This cascading design penalizes intensity errors only after correct expression recognition. Table 6 reports state-of-the-art results. The validated facial captioner is then used to generate expression instructions for real-world image pairs. Figure 10 illustrates several training samples from HFL-150K.



Make her facial expression happy slightly

Make his facial expression happy slightly and turn his head 30° right

Make her facial expression happy strongly

Make his facial expression sad slightly and turn his head 30° left

Make his facial expression sad slightly

Make her facial expression sad strongly

Make her facial expression surprised strongly and turn her head 30° down

Make his facial expression surprised

Make his facial expression surprised slightly and turn his head 30° right

Make her facial expression angry

Make her facial expression angry strongly

Make his facial expression angry strongly

Make her facial expression disgusted slightly and turn her head 30° right

Make her facial expression disgusted strongly

Make her facial expression disgusted slightly

Turn his head 30° left

Turn his head 30° right

Turn his head 60° left

Figure 10: Examples from HFL-150K. Each triplet shows a source image, its instruction, and the corresponding target edit.

B.3 Diversity Statistics

Following standard practice in population studies (Buolamwini & Gebru, 2018), we estimate age, gender, and skin tone for HFL-150K. The dataset contains 34.2% subjects aged 40 or above, 48.3% female and 51.7% male, and 23.1% dark-skinned individuals, with the remaining subjects distributed across medium skin tones at 45.7% and light skin tones at 31.2%. In-

Table 6: Comparison of captioning performance across expression granularity levels.

Methods	Accuracy
Qwen2.5-VL-7B	46.5
Qwen2.5-VL-72B	56.9
Ours	67.3

Table 7: Scaling analysis of HFL-150K.

Training size	SC ↑	VQ ↑	NA ↑	IP ↑
20K	0.784	0.728	0.749	0.571
50K	0.813	0.735	0.767	0.593
100K	0.828	0.746	0.782	0.617
150K	0.832	0.749	0.805	0.634

Table 8: User study results.

Methods	SC	VQ	IP	Overall
Step1X-Edit†	0.728	0.810	0.746	0.761
Flux.1-Kontext†	0.712	0.824	0.774	0.770
LaTo	0.766	0.837	0.829	0.811

tuitively, this indicates that the dataset is not overly concentrated on a narrow demographic, but instead offers a reasonably balanced coverage of different groups. We further observe that these proportions remain stable when we focus on challenging cases such as extreme head poses and strong facial expressions, suggesting that the demographic diversity is preserved even in hard scenarios.

C MORE EXPERIMENTS

C.1 SCALING ANALYSIS OF HFL-150K

To assess HFL-150K's scaling behavior, we train LaTo on stratified subsets of 20K, 50K, 100K, and 150K samples while preserving class distribution. All models are trained from scratch under identical settings. Table 7 shows clear scaling trends: instruction adherence and visual quality improve rapidly at moderate scales (50K, 100K), whereas identity consistency and photorealism depend more strongly on larger, higher-quality data, with the full 150K set yielding the most robust gains. These results indicate that increasing data volume improves geometric modeling, which in turn strengthens identity preservation and natural appearance. Systematic expansion of the dataset via our pipeline is expected to further advance large-scale facial editing.

C.2 USER STUDY

We conduct a user study on all 1,110 cases across the three benchmarks and evaluate Semantic Consistency, Visual Quality, and Identity Preservation using a 5-point Likert scale for fine-grained assessment. The detailed questionnaires are provided in Section D.3. For a fair comparison, we evaluate Flux.1-Kontext† and Step1X-Edit†, both fine-tuned on the HFL-150K training set, and obtain 21 valid participant responses. As shown in Table 8, LaTo achieves an overall rating that is about 4.1% higher than Flux.1-Kontext†, indicating that users consistently perceive LaTo's edits as better aligned with the target semantics, visually more plausible, and more faithful to the original identity, and therefore more suitable for fine-grained facial editing in practical applications.

C.3 EXTENSION TO VIDEO-BASED FACIAL ANIMATION

We extend both the Landmark Tokenizer and the location-mapping positional encoding to facial animation, enabling controllable video-avatar generation. The key differences from LaTo are as follows. For data, we construct a 16k high-quality single-portrait dataset from OpenHumanVid using similar filtering criteria, with each video annotated by landmark trajectories and descriptive captions. For the tokenizer, we generalize it to 3D with a temporal length of 81 frames. For fusion, we adopt Wan2.2-TI2V-5B (Wan et al., 2025) as the base model and introduce an additional cross-attention layer after the text-to-video attention. During training, we freeze all original weights and only train this new cross-attention, which makes the adaptation efficient and avoids disrupting the pretrained text-video alignment. The visual results in Figure 11 show that LaTo maintains robust identity preservation when editing with cross-identity reference face videos, demonstrating strong scalability. However, the current system is reference-video driven, and further work is needed to support interactive control

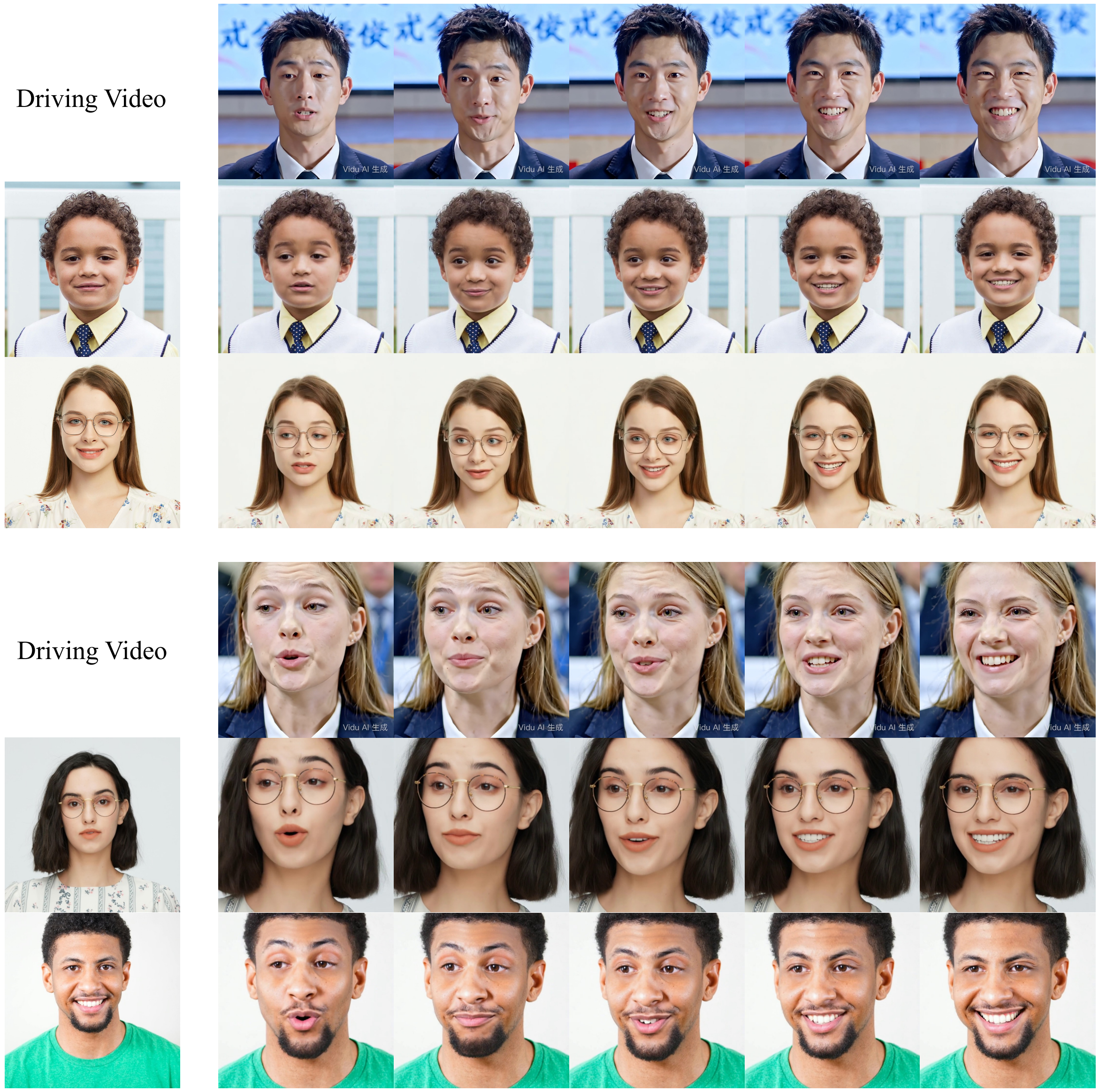


Figure 11: Examples of facial animation generated by LaTo. Given a cross-identity driving video, LaTo can also achieve identity–geometry decoupling in the video domain, preserving the target identity while following the facial motions of the driving video.

similar to LaTo. These include developing a reliable landmark predictor for generating landmark sequences and identifying a more optimal fusion strategy, as the current additional cross-attention may be suboptimal.

C.4 Extension to full-body editing

For other structured domains, we further extend LaTo's design to full-body editing. The corresponding visualizations are presented in Figure 12. We observe that cross-identity input poses a greater challenge for full-body editing than for faces, as body shapes exhibit substantially higher inter-subject variability. The baseline's pixel-wise alignment tends to "paste" the body structure from the conditional image, leading to pronounced identity distortions under cross-identity settings. In contrast, LaTo preserves the source subject's overall body structure while still adhering to the target pose condition, yielding more stable and identity-consistent editing. A promising research direction is to jointly model body and facial tokenizers to enable more fine-grained, identity-preserving human image editing.

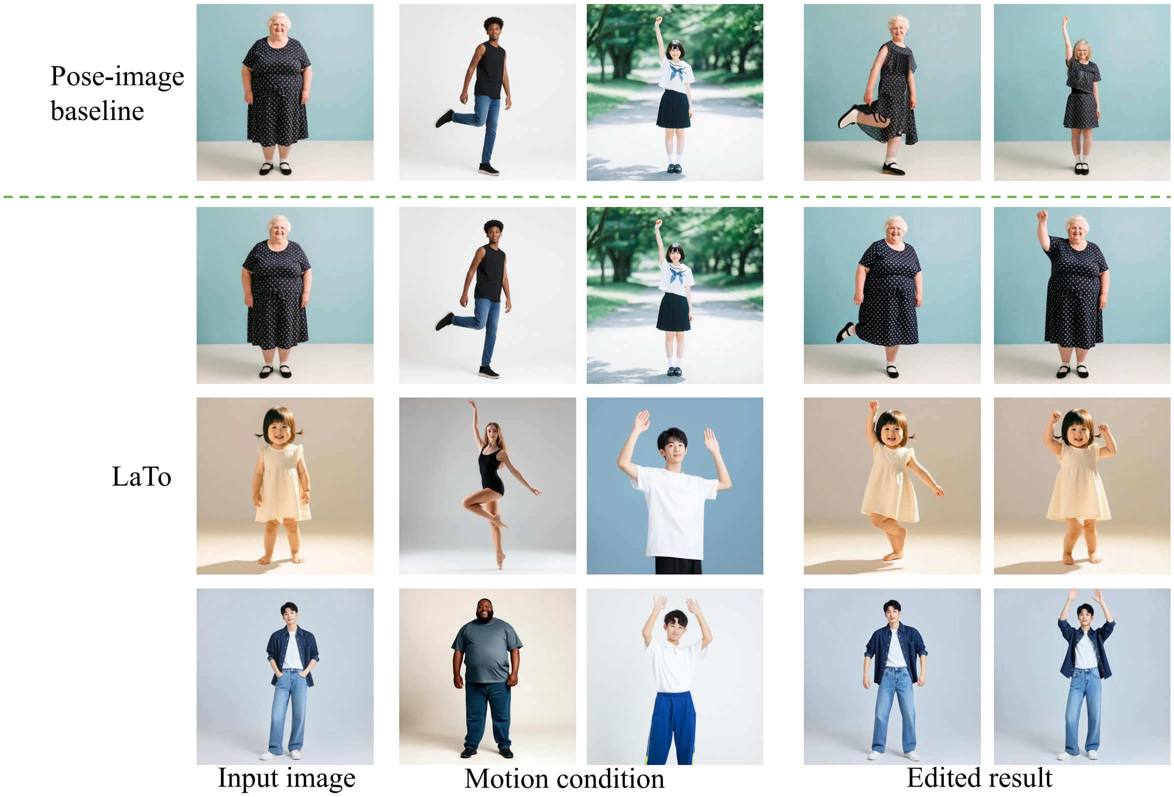


Figure 12: Comparison results of full-body editing between the pose–image baseline and LaTo.

C.5 Comparison of Landmark Predictors

Figure 13 compares qualitative outputs across landmark predictors. Qwen2.5-VL produces unstable, low-magnitude updates. It under-responds to instructions that demand large pose or expression changes, yielding displacements clustered near the source landmarks, and it often ignores low-contrast regions, leaving jawline and cheek shifts static. Gemini 2.5 Pro better follows the requested direction but frequently distorts facial geometry. Common failure cases include global scaling that alters inter-ocular distance, jawline widening or compression, and asymmetric eyebrow motions not supported by the instruction. These behaviors suggest that general-purpose VLMs lack task-specific supervision for landmark reasoning. In contrast, our landmark predictor, fine-tuned for structured chain-of-thought landmark reasoning, accurately captures pose- and expression-induced landmark shifts while preserving facial shape, establishes a stable mapping between anatomical features and fine-grained linguistic instructions, and enables downstream edits that preserve identity while faithfully realizing the requested pose and expression changes.

C.6 Design Choices for Unconditional Landmark Tokens

We investigate the trade-off between landmark alignment and instruction fidelity in the unconditional branch of classifier-free guidance. Two strategies are compared: a zero landmark list mapped to the tokenizer's embedding space, and learnable unconditional landmark tokens. As shown in Figure 14, the zero-embedding strategy exhibits unstable optimization with divergent losses and strong sensitivity to guidance scales. Zero vectors fail to represent an unconstrained landmark state, which conflicts with the dynamics of facial geometry. This mismatch leads to two characteristic failures: image-wide artifacts and weak alignment with the target landmarks and instruction (Figure 15). In contrast, learnable unconditional tokens produce stable training, encode the unconstrained state explicitly, and integrate landmark guidance with instruction-based editing to achieve coherent cross-modal alignment.

C.7 Rectified Identity Preservation Score

We find that identity evaluation can be biased when a model avoids making the requested edit. Even advanced metrics such as ArcFace or large VLMs report high similarity despite clear instruction violations. In Figure 16, the instruction "make his facial expression scared" is ignored by Step1X-Edit. The output shows minimal change($\Delta SSIM = 0.05$) while ArcFace

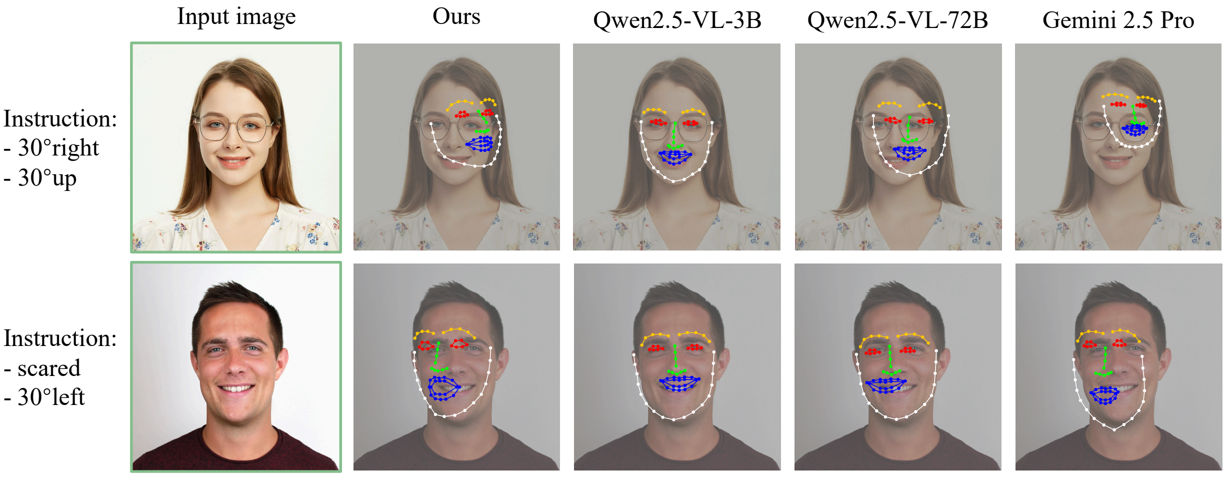


Figure 13: Qualitative comparison of our landmark predictor with existing VLMs.

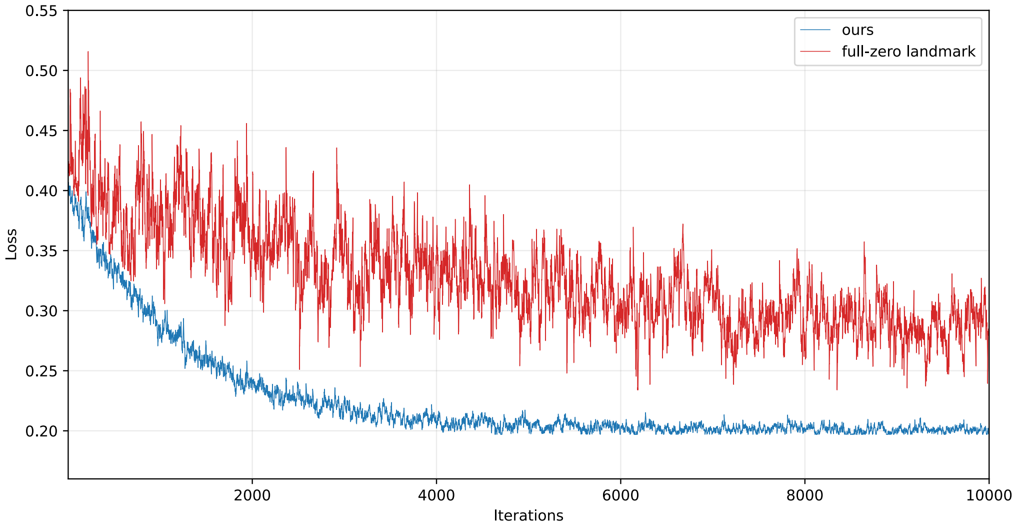


Figure 14: Training loss across iterations for different unconditional landmark token implementations.

remains high at 0.98, which fails to reflect the editing intent. To address this limitation, we propose a rectified IP score that couples identity similarity with transformation magnitude and instruction compliance. As outlined in Algorithm 1, the Qwen2.5-VL first identifies facial regions in the source image and predicts the expected editing amplitude φ_{ins} from the instruction. We then measure the realized change φ_{real} using SSIM between the source and edited images and compute a discrepancy penalty p with a factor α = 2. The rectified score s_{rip} combines the identity score s_{arc} with this penalty. For the example in Figure 16: Step1X-Edit yields $\varphi_{ins} = 0.257, \varphi_{real} = 0.05, s_{arc} = 0.984 \Rightarrow p = 0.648, s_{rip} = 0.336$, while LaTo achieves $\varphi_{ins} = 0.257, \varphi_{real} = 0.341, s_{arc} = 0.759 \Rightarrow p = 0.065, s_{rip} = 0.694$.

We validate the rectified IP score with a user study. We generate win–lose pairs for Step1X-Edit and LaTo using the original IP metric and the rectified IP metric. Participants judge whether the score difference aligns with their perception under the given instruction. The rectified IP score aligns with human preference with a 7% error rate, which substantially outperforms ArcFace at 46%.

C.8 Extended Qualitative Comparisons

Figure 18 presents additional qualitative results under diverse instructions. LaTo follows the specified edits while preserving identity and facial geometry. Baselines often produce artifacts, unnatural appearance, or landmark drift. These results further demonstrate the effectiveness of our design for controllable face editing.

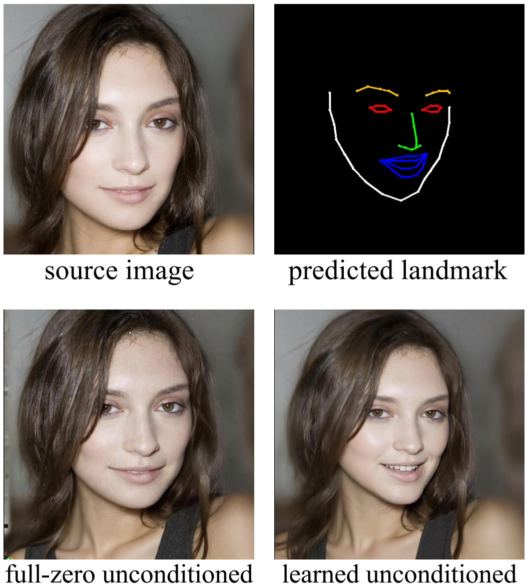


Figure 15: Qualitative visualization of generated outputs using different unconditional landmark token strategies.

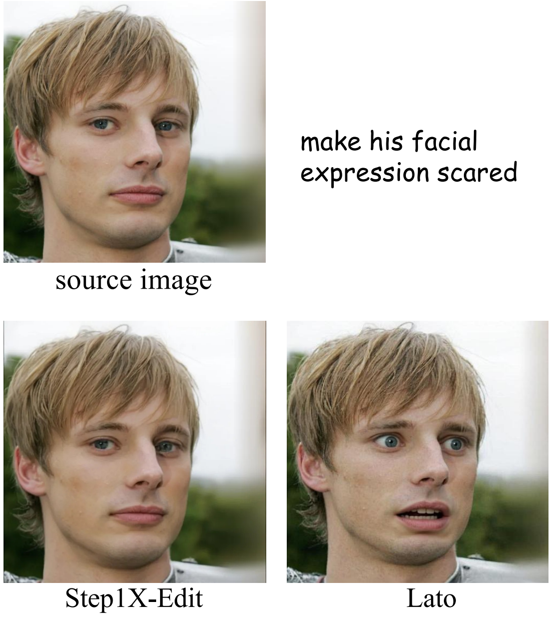


Figure 16: Example of minor edits that contradict the editing instructions.

Algorithm 1 Rectified Identity Preservation Score Calculation

```
Require: Source image $I_s$, edited image $I_t$, editing instruction $T$
Ensure: Final score $s_{rip}$
$s_{arc} \leftarrow \text{ArcFace}(I_s, I_t)$                      $s_{arc} \in [0,1]$
$\varphi_{ins} \leftarrow \text{GetExpectedMagnitude}(I_s, T)$      $\varphi_{ins} \in [0,1]$ via QwenVL prediction
$\varphi_{real} \leftarrow \text{SSIMDiff}(I_s, I_t)$               $\varphi_{real} \in [0,1]$
$\alpha \leftarrow 2, \epsilon \leftarrow 1e-5$                      Hyperparameters
$$p \leftarrow (\frac{\varphi_{ins} - \varphi_{real}}{\varphi_{ins} + \epsilon})^{\alpha}$$
$s_{rip} \leftarrow \max(0, s_{arc} - p)$
return $s_{rip}$
```

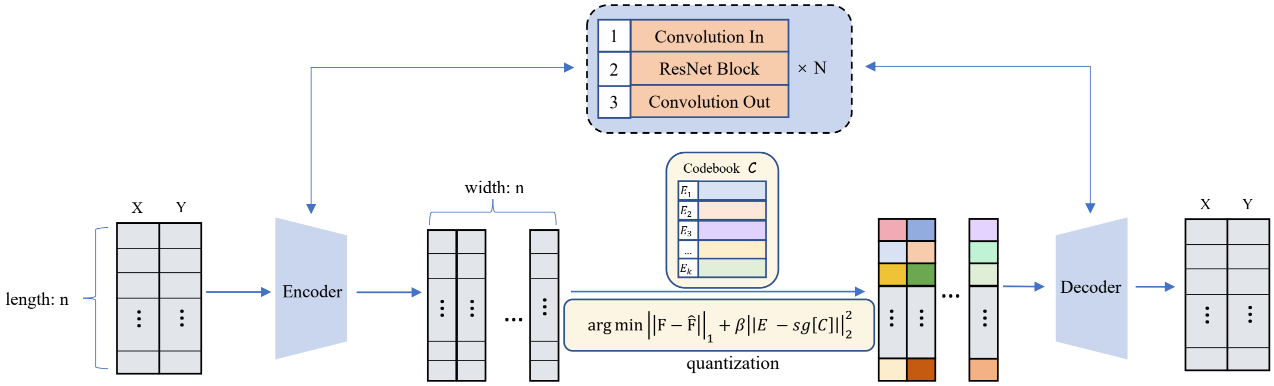


Figure 17: The architecture of the landmark tokenizer.

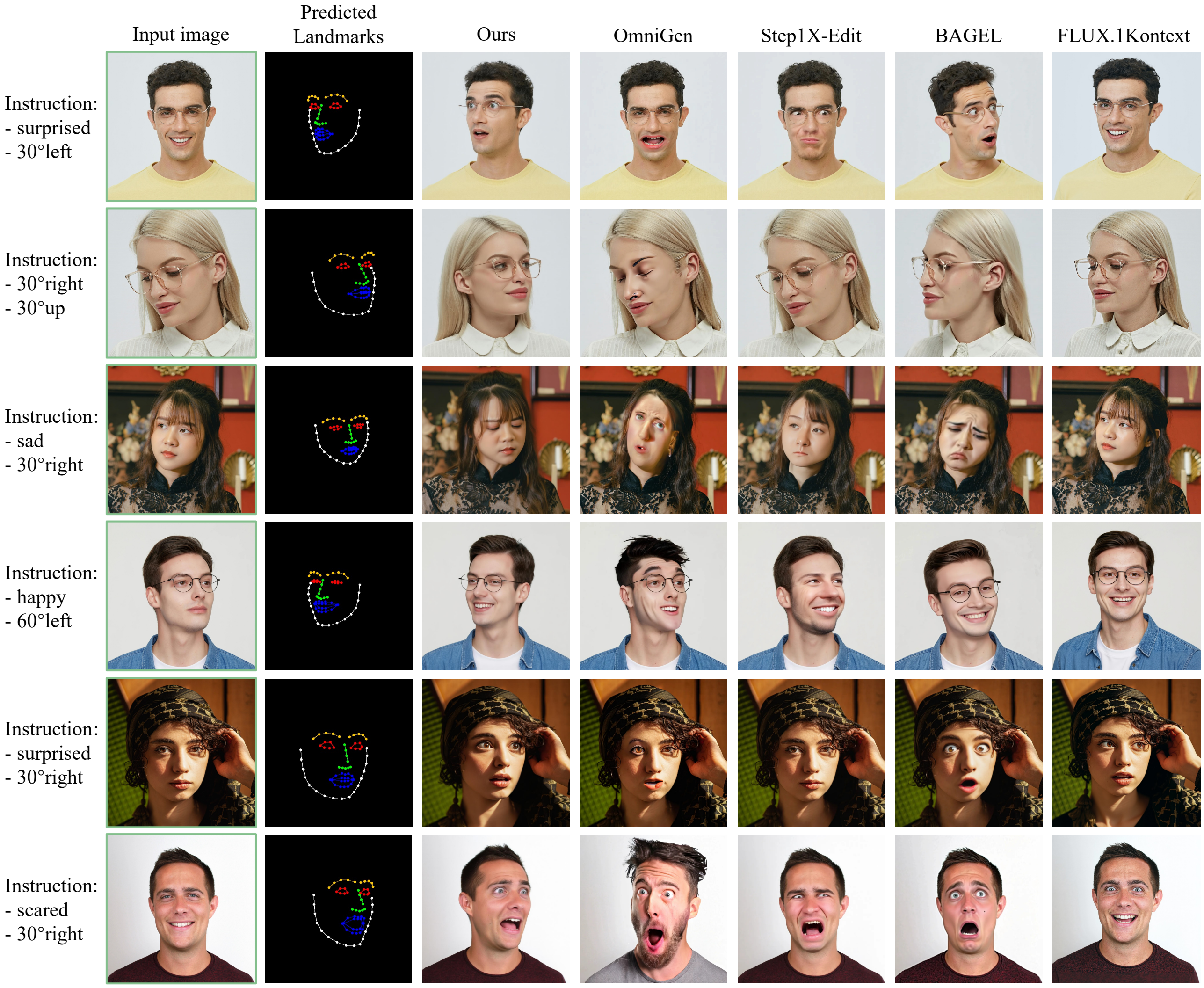


Figure 18: Additional comparisons with state-of-the-art editing methods.

D Additional Methodological Details

D.1 Landmark Tokenizer

Figure 17 shows the detailed architecture of the landmark tokenizer. To quantify codebook diversity, we computed cosine similarities for all pairs of code vectors and report aggregate statistics: mean ≈ 0.007 and standard deviation ≈ 0.019 on the test set. The near-zero mean and low dispersion indicate that the code vectors are close to orthogonal, confirming that the tokenizer induces a discrete embedding space with strong vector distinctiveness and minimal redundancy.

We also evaluated codebook sizes of 1,024, 8,192, and 16,384, as well as a no-quantization ablation. Table 9 reports landmark reconstruction performance on HFL-150K. For PSNR evaluation, we render the landmarks into images. Removing quantization forces the tokenizer to operate on continuous and inconsistent facial features, which degrades performance. This confirms that discrete and general landmark tokens provide more stable geometric representations and improve generalization. Although larger codebooks offer higher nominal capacity, the 16,384-entry codebook resulted in unstable optimization and significant codebook underutilization, which reduced its effective capacity. The 8,192-entry codebook achieved the best overall performance. We anticipate further gains through improved codebook utilization.

D.2 Structured Chain-of-Reasoning for the Landmark Predictor

D.2 presents the prompt used to elicit a structured chain of reasoning conditioned on the source image, the instruction, and the target landmark. The template enforces a machine-parsable schema that includes:

Table 9: Ablation of codebook size.

Codebook Size	Utilization (%)	L1 Distance	PSNR
no-quantization	-	7.04	19.11
1024	81.42	5.32	20.47
8192	70.39	3.67	21.33
16384	44.75	4.51	20.61

Facial Landmark Prediction Chain-of-Thought Prompt

`### Role and Task`

You are an expert in facial geometry, computer vision, and human facial expression analysis. Your task is to generate a detailed, explicit, step-by-step Chain-of-Thought (CoT) reasoning process that shows exactly how to derive the edited facial landmark coordinates (target output) by reasoning from the following four inputs:

1. The input image of a human face.
2. The original facial landmarks of the face (given as precise coordinate arrays, typically grouped by facial regions such as JAW/BROWS, NOSE, EYES, MOUTH).
3. The natural language editing instruction describing how the face should be transformed (e.g., change in expression, head pose, or other features).
4. The edited facial landmarks (target result coordinates after applying the transformation).

Your output is NOT to be a direct mapping or a simple coordinate difference. Instead, it must be a structured internal reasoning narrative that reflects the plausible human-expert thought process in getting from the original state (using 1, 2, 3) to the edited state (4). This reasoning should:

- Begin with an **Initial State Analysis** using the input image and original landmarks to describe the current facial pose, expression, and alignment.
- Include anatomical and geometric interpretation of the original landmarks and how they correspond to key facial features.
- **Decompose the editing instruction** into precise, separate transformations (e.g., rigid 3D head rotation, non-rigid facial muscle changes).
- Analyze **which landmark groups** are affected by each transformation and how (both direction and magnitude of expected movement in image coordinate space).
- Consider both **global rigid transformations** (e.g., head rotation, translation, scaling effects due to perspective) and **local non-rigid deformations** (e.g., eye widening, eyebrow lifting, mouth shape change).
- Provide quantitative or semi-quantitative estimates of coordinate shifts (in X/Y pixels) that are consistent with the original and final coordinates.
- Explain how perspective and foreshortening influence the new positions after a head turn.
- Show how expression changes (like smiling → scared) affect local geometry, including muscle tension, lip curvature, jaw drop, eyelid movement, eyebrow displacement.
- Maintain **identity coherence** in the resulting shape: distances and proportions in rigid facial structures should remain consistent with physical reality.
- Explicitly connect initial landmark positions (from 2) to final ones (from 4) through the reasoning steps.

The final Chain-of-Thought should be **comprehensive, logically progressive, and reflect expert-level spatial reasoning**—not just descriptions of differences. Use a numbered step-by-step format (e.g., Step 1, Step 2...) and include both qualitative anatomical analysis and approximate numerical transformation where applicable. Always explain *why* each change happens given the transformation described in the instruction, and ensure the reasoning is consistent with realistic facial kinematics. Output the CoT reasoning directly, without any irrelevant descriptions.

`### Inputs for your reasoning will be presented in the following format:`

1. {Input image}
2. {Original Landmarks JSON}
3. {Editing Instruction}

4. {Edited Landmarks JSON}

`### Chain-of-Thought Reasoning you should produce:`

Step 1: Initial State Analysis.

- Analyze the Visual Reference Image to understand the face's starting condition.
- Current Pose:
- Current Expression:
- Correlate the visual features with the provided Initial Normalized Landmark Coordinates to build a mental model of the face.

Step 2: Instruction Decomposition and Kinematic Analysis.

- Break down the instruction <Your Input Editing Instruction> into primary anatomical movements.
- Primary Action(s):
- Key Facial Parts Affected:
- Kinematic Chain Reasoning: (Describe how these parts move together as a 2D projection within the 512x512 normalized coordinate space.)

Step 3: Quantitative Transformation Estimation (in the 512x512 space).

- Translate the qualitative reasoning into quantitative coordinate shifts for landmark groups. All estimations refer to the 512x512 normalized canvas.
- Example Transformation for 'Turn right and smile':
  - `jawline (points 0-16)`: "Left points (0-8) will shift right (positive X) by about 25–40 units. Right points (9-16) will also shift right, but by a smaller amount, maybe 10–20 units."
  - `nose (points 27-35)`: "All points will shift right (positive X). Point 33 (tip) will move the most, maybe 30 units. Point 27 (root) will move the least, maybe 8 units."
  - `mouth (points 48-67)`: "Points 48 and 54 will move up (negative Y) by 15 units and horizontally apart by 10 units each. The center of the upper lip (point 51) will rise (negative Y) by 8 units."
- Identity Constraint: "All transformations must be cohesive. The relative distances within rigid groups (like the nose bridge) should be largely preserved, while deformable areas (like the mouth) change shape according to the expression. The overall transformation should look like the same person."

`### Input:`

1. [Input image]
2. Original Landmarks JSON:

```
{"JAW/BROWS": [[160, 198], [160, 226], [164, 247], [171, 267], [178, 291], [191, 312],
     [202, 322], [219, 336], [250, 346], [277, 339], [298, 326], [315, 315], [329,
    295], [336, 271], [339, 247], [346, 226], [346, 198], [174, 174], [184, 167], [198,
     167], [212, 167], [226, 174], [274, 174], [284, 171], [298, 167], [315, 171],
    [325, 178]], "NOSE": [[250, 202], [246, 222], [246, 236], [246, 250], [233, 257],
    [236, 260], [246, 264], [257, 260], [267, 257]], "EYES": [[191, 198], [202, 195],
    [212, 195], [226, 202], [215, 205], [202, 202], [274, 202], [284, 198], [298, 198],
     [305, 202], [298, 205], [284, 205]], "MOUTH": [[212, 281], [222, 278], [239, 278],
     [250, 278], [257, 278], [274, 278], [288, 284], [274, 298], [260, 305], [246,
    308], [236, 305], [222, 298], [215, 281], [236, 284], [250, 284], [260, 284], [288,
     284], [260, 291], [246, 295], [236, 291]]}
```

3. Editing Instruction: turn his/her head 30 degrees to the right and 30 degrees up
4. Edited Landmarks JSON:

```
{"JAW/BROWS": [[192, 198], [197, 220], [202, 242], [211, 263], [226, 280], [248, 291],
     [275, 298], [301, 303], [324, 303], [341, 298], [350, 285], [357, 267], [362,
    249], [364, 231], [365, 212], [361, 195], [355, 179], [233, 164], [243, 151], [260,
     143], [277, 142], [294, 147], [317, 146], [327, 139], [337, 138], [346, 141],
    [351, 151]], "NOSE": [[310, 167], [316, 176], [323, 185], [330, 195], [305, 217],
    [315, 217], [324, 218], [329, 216], [334, 214]], "EYES": [[255, 179], [265, 170],
    [275, 169], [283, 176], [275, 179], [265, 180], [320, 174], [329, 166], [338, 165],
     [344, 173], [338, 175], [329, 175]], "MOUTH": [[281, 249], [298, 238], [315, 232],
    [325, 234], [332, 231], [340, 234], [342, 245], [341, 255], [335, 262], [326,
  264], [316, 264], [299, 260], [286, 249], [315, 242], [325, 241], [332, 241], [339,
    245], [333, 250], [326, 252], [316, 253]]}


### Output:
```

D.3 Face Editing Questionnaire

Thank you for completing the following rating of image editing results. Please spend about 15 seconds on each image, consider the source image and the editing instruction, and rate the result based on the following dimensions:

1. Semantic Consistency
Does it match the text prompt (e.g., "Smiling slightly and turn her head left 30 degrees")?
1: No match at all.
2: Very weak or incorrect interpretation (e.g., wrong expression or pose direction).
3: Partial match — correct type but inaccurate intensity or rotation degree.
4: Mostly accurate — small deviation in expression strength or head angle.
5: Perfectly matches the prompt in both expression and head pose.

2. Visual Quality
How photorealistic and artifact-free is the generated face?
1: Severely distorted or unrealistic — obvious artifacts, broken facial structure, or non-human appearance.
2: Low quality — blurry, noisy, or with noticeable distortions (e.g., warped eyes, asymmetric features).
3: Acceptable but flawed — overall plausible, but has mild blurriness, texture glitches, or minor asymmetry.
4: High quality — sharp and realistic, with only subtle imperfections hard to notice at a glance.
5: Excellent quality — indistinguishable from a real photo; no visible artifacts or distortions.

3. Identity Preservation
Is it the same person as the source?
1: Completely different person.
2: Unlikely the same — key facial features (e.g., eye spacing, nose shape, jawline) don't match.
3: Uncertain — could be the same or different.
4: Likely the same — minor structural differences, but core identity preserved.
5: Definitely the same person — identity is unmistakable.

E Limitations

Though LaTo is effective for fine-grained editing and identity preservation, the complex architecture of the advanced base model limits its ability to support efficient on-the-fly processing, so we rely on future advances in accelerating such base models by the community. Moreover, we observe that for input images with extreme head poses caused by heavy occlusion, landmark estimation becomes less reliable, leading to slight degradation in identity preservation. We plan to incorporate 3D-aware priors and richer occlusion annotations to mitigate these limitations.