

CLIP-FTI: Fine-Grained Face Template Inversion via CLIP-Driven Attribute Conditioning

Longchen Dai, Zixuan Shen*, Zhiheng Zhou, Peipeng Yu, Zhihua Xia*

College of Cyber Security, Jinan University
 {LongchenDai, jnuzzh}@stu2023.jnu.edu.cn, {m15195901120, ypp865, xia_zhihua}@163.com

Abstract

Face recognition systems store face templates for efficient matching. Once leaked, these templates pose a threat: inverting them can yield photorealistic surrogates that compromise privacy and enable impersonation. Although existing research has achieved relatively realistic face template inversion, the reconstructed facial images exhibit over-smoothed facial-part attributes (eyes, nose, mouth) and limited transferability. To address this problem, we present CLIP-FTI, a CLIP-driven fine-grained attribute conditioning framework for face template inversion. Our core idea is to use the CLIP model to obtain the semantic embeddings of facial features, in order to realize the reconstruction of specific facial feature attributes. Specifically, facial feature attribute embeddings extracted from CLIP are fused with the leaked template via a cross-modal feature interaction network and projected into the intermediate latent space of a pretrained StyleGAN. The StyleGAN generator then synthesizes face images with the same identity as the templates but with more fine-grained facial feature attributes. Experiments across multiple face recognition backbones and datasets show that our reconstructions (i) *achieve higher identification accuracy and attribute similarity*, (ii) *recover sharper component-level attribute semantics*, and (iii) *improve cross-model attack transferability compared to prior reconstruction attacks*. To the best of our knowledge, ours is the first method to use additional information besides the face template attack to realize face template inversion and obtains SOTA results.

Introduction

Face recognition (FR) systems ubiquitously store facial templates (i.e., compact deep embeddings extracted during enrollment) for efficient verification and identification across applications such as smartphone unlocking and border control. Once leaked, these templates can be used to attack FR systems: face template inversion (FTI) attacks can reconstruct photorealistic facial surrogates that simultaneously (i) compromise user privacy (exposing soft biometric attributes such as age, gender, ethnicity, or skin tone) and (ii) enable downstream impersonation attacks. Unlike most attacks on FR systems that threaten the security of the system (Biggio et al. 2015; Deb, Zhang, and Jain 2020; Fei et al. 2022)

and biometric privacy (Galbally, Marcel, and Fierrez 2014; Galbally et al. 2010; Hadid et al. 2015; Marcel, Fierrez, and Evans 2023; Marcel et al. 2019; Fei et al. 2020). FTI directly undermines both privacy and security by recovering an identity’s image from its leaked template and thus requires further study.

Early FTI attacks attempt employed custom encoder-decoder networks (Mai et al. 2018) that directly mapped templates back to pixel space, but were constrained by limited generative capacity and low visual fidelity. Recent methods leverage powerful pretrained generative adversarial networks (GANs) (Otroski, Shahreza, and Marcel 2023), first projecting facial templates into an intermediate latent space, and then synthesizing higher-resolution face images. While these GAN-assisted approaches markedly improve realism, they predominantly rely on a single leaked face template. As a result, reconstructions often exhibit over-smoothed facial feature attributes (periocular shape, nasal ridge, lip contours).

Text-conditioned image synthesis provides a new approach for the fine-grained recovery of facial template attributes. Relevant studies shows that semantic descriptors can steer generative models toward attribute-faithful refinements without eroding global structure. Inspired by this, we hypothesize that augmenting leaked templates with semantic attribute embeddings can (i) enrich reconstruction details and (ii) elevate cross-model transferability.

We introduce **CLIP-FTI** (Face Template Inversion via Fine-Grained Attribute Conditioning), a CLIP-driven face template inversion framework. Given a leaked facial template t , we derive facial semantic attribute embeddings using pretrained CLIP model by matching the template’s paired image against a curated bank of textual prompts that describe fine facial feature attributes (e.g., eye-shape variants, nasal bridge traits, lip fullness descriptors). These attribute embeddings are aggregated into a single representation s . Then we use two learned mapping sub-modules of CLIP-FTI: (1) a *template* \rightarrow *attribute alignment* (TAA) adapter that predicts the CLIP-based attribute embedding s from the leaked template t ; and (2) a *fusion* \rightarrow *latent projector* that first fuses t and s into a joint code f (via cross-modal attention) and then maps f to a StyleGAN latent code \mathcal{W} . The frozen StyleGAN generator subsequently synthesizes an identity-consistent, attribute-faithful face image. By conditioning on

*Zhihua Xia and Zixuan Shen are co-corresponding authors.
 Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

CLIP-derived facial attribute semantic embedding in a single forward pass, CLIP-FTI compensates for detail defects of prior template-only inversions, yielding higher identification accuracy, sharper attribute similarity, and stronger cross-model transferability.

To elaborate on the contributions of this paper, we list them hereunder:

- To our best knowledge, we make the first attempt to explore face template inversion using CLIP-driven semantic attribute embeddings, mitigating over-smoothed facial feature attribute details prevalent in prior method. Specifically, we design a cross-modal feature interaction network that fuses leaked templates with CLIP attribute embeddings and maps the fused representation into StyleGAN’s intermediate latent space, explicitly addressing the identity-detail trade-off.
- We introduce a supervised mapping from facial templates to the corresponding attribute embeddings, enhancing semantic compatibility and improving downstream reconstruction fidelity.
- Experiments across multiple FR backbones and datasets show consistent gains in attribute similarity, identification accuracy, and cross-model attack transferability over state-of-the-art inversion attacks.

Related Work & Preliminary

Face Template Inversion

The early facial template inversion attacks typically trained bespoke encoder-decoder networks that map a leaked template to a low-resolution image without any generative prior. Mai *et al.* (Mai et al. 2018) and Cole *et al.* (Cole et al. 2017) jointly minimised pixel and perceptual loss, but the absence of a strong generator led to blurry faces and poor high-frequency detail, even in white-box settings. With the advent of StyleGAN, most subsequent works learned a *global* mapping from a template to the generator latent space. Dong *et al.* (Dong et al. 2021) and Otroshi *et al.* (Otroshi Shahreza and Marcel 2023) regress the template to StyleGAN2/3 latent space via an MLP, whereas Shahreza *et al.* (Shahreza and Marcel 2024, 2023) exploit synthetic pairs or a NeRF-based EG3D backbone. A complementary line searches the latent space per target—e.g. simulated annealing (Vendrow and Vendrow 2021) or genetic algorithms (Dong et al. 2023). These black-box attacks can recover identity but require hundreds of forward passes per image, making them computationally prohibitive. Although these methods inherit the visual quality of modern generators, relying on a single latent vector produces over-smoothed facial feature attributes (eyes, nose, and mouth), limiting both realism and attack transferability.

Vision–Language Models and CLIP

Contrastive Language–Image Pre-training (CLIP) (Radford et al. 2021) aligns images and text in a shared embedding space by learning from millions of (image, caption) pairs. Its strong zero-shot transfer has spurred a series of generative applications, including prompt-based image editing (StyleCLIP (Patashnik et al. 2021)), domain

Method	Year	Gen./Res.	FG	1-shot
Cole <i>et al.</i>	2017	CNN / 128	✗	✓
Mai <i>et al.</i>	2018	CNN / 128	✗	✓
FaceID-GAN	2021	SG2 / 512	✗	✓
Otroshi <i>et al.</i>	2023	SG3 / 1024	✗	✓
Vendrow <i>et al.</i>	2021	SG2 / 1024	✓	✗
Dong <i>et al.</i>	2023	SG2 / 1024	✓	✗
CLIP-FTI (ours)	2025	SG3 / 1024	✓	✓

Table 1: Comparison with prior facial template inversion attacks. "FG" = recovers fine-grained attribute details; "1-shot" = single forward pass at inference; "SG2/3" = StyleGAN2/3.

adaptation (StyleGAN-NADA (Gal et al. 2022)), and text-guided 3D synthesis (DreamFusion (Poole et al. 2022)). CLIP2StyleGAN (Abdal et al. 2021) uses CLIP to discover and name disentangled semantic directions and then edits faces by moving StyleGAN latents along those directions; Arc2Face (Paraperas Papantoniou et al. 2024) replaces text conditioning with an ArcFace ID embedding to synthesize high-quality, ID-consistent faces with a diffusion backbone.

In contrast, our goal is a targeted face template inversion (FTI) attack. Unlike general CLIP-guided editing or ID-consistent synthesis, we use CLIP to extract fine-grained facial-attribute embeddings (eyes, nose, mouth) and fuse them with the leaked template in a single pass. This guidance enforces text-semantic fidelity and yields high-resolution reconstructions that retain local details while fooling the target FR model. Prior template attacks often produce over-smoothed faces; our attribute-guided design mitigates this and improves attack efficacy. A concise comparison appears in Table 1.

Preliminary

StyleGAN. StyleGAN (Otroshi Shahreza and Marcel 2023; Karras, Laine, and Aila 2019; Karras et al. 2020) employs a mapping network $f : \mathcal{Z} \rightarrow \mathcal{W}$ that transforms an input noise vector $z \sim P_Z$ (e.g., Gaussian) into an intermediate latent code $w = f(z)$. This disentangled latent space \mathcal{W} is fed (via per-layer affine transformations) into the synthesis network to modulate convolutional weights (style modulation). Hierarchical semantics emerge: early layers control coarse pose and face shape, middle layers affect facial components (eyes, nose, mouth), and later layers refine texture and color micro-details.

CLIP. CLIP (Radford et al. 2021) jointly trains an image encoder $E_I(\cdot)$ and a text encoder $E_T(\cdot)$ on large-scale image-text pairs using a symmetric contrastive loss. Given a mini-batch $\{(x_i, d_i)\}_{i=1}^B$, where x_i is an image and d_i is its paired textual description, the encoded features are ℓ_2 -normalized and used to form a temperature-scaled cosine-similarity matrix. Cross-entropy then encourages each x_i to match only its corresponding d_i (and vice versa) while pushing apart mismatches. This yields a shared embedding space in which semantically related images and texts lie nearby. Empirically, CLIP captures a broad range of visual concepts, including fine facial feature attributes (e.g., eye shape, brow thick-

ness, nose-bridge height, lip fullness, skin-tone variations), despite not being explicitly trained for biometrics.

Method

In this section, we formalize the face template inversion (FTI) threat scenario, and introduce our CLIP-driven fine-grained attribute conditioning model **CLIP-FTI**.

Attack scenario. We consider a FTI attack against a FR system based on the following scenario:

- **Adversary’s goal:** The adversary seeks to reconstruct a face image from a leaked facial template, and use the reconstructed face image to enter the target FR system.
- **Adversary’s knowledge:** The adversary is assumed has access to a facial template, which was leaked from the database of the target’s FR system. But lacks knowledge of the internal structure of the FR model (blackbox) and can only interact with the model through external interfaces (such as SDKs, APIs, etc.) for any face image.
- **Adversary’s capability:** The adversary is able to utilize the reconstructed face image to access the target face recognition system. For simplicity, it is assumed that the adversary can inject the reconstructed face image as a query into the target system.
- **Adversary’s approach:** The adversary can train a face reconstruction model to invert facial templates and reconstruct underlying face images. Then, the adversary can use reconstructed face images to inject as a query to the target FR system, to enter that system.

CLIP-FTI. To perform fine-grained face template inversion, we introduce **CLIP-FTI**, a CLIP-driven attribute conditioning framework for face template inversion. During *training* stage, we assume access to face images and their corresponding recognition templates (extracted by a surrogate FR model) so that we can (i) extract CLIP-based semantic attribute embeddings and (ii) learn two mappings: a template→attribute alignment (TAA) adapter and a fusion→latent projector (M_{FLP}). During the *attack* (inference) stage, only a leaked template t is available; the trained TAA adapter predicts a semantic attribute embedding \hat{s} , which together with a noise vector z is fused to produce a latent code $\hat{w} \in \mathcal{W}$; the frozen StyleGAN3 generator then synthesizes the reconstructed face $\hat{I} = G(\hat{w})$. Figure 1 summarizes the pipeline.

TAA Adapter

Facial feature attribute prompt matching. To enable more detailed semantic alignment, we divide the face into multiple regions—such as eyes, nose, and mouth—and predefine a set of textual descriptions for each region:

$$\begin{aligned} D_{\text{eyes}} &= \{d_{\text{eyes}}^1, d_{\text{eyes}}^2, \dots, d_{\text{eyes}}^n\}, \\ D_{\text{nose}} &= \{d_{\text{nose}}^1, d_{\text{nose}}^2, \dots, d_{\text{nose}}^n\}, \\ &\dots \\ D_{\text{mouth}} &= \{d_{\text{mouth}}^1, d_{\text{mouth}}^2, \dots, d_{\text{mouth}}^n\}. \end{aligned}$$

Each description d_i is encoded into a textual feature vector v_i via the text encoder $E_T(\cdot)$ of CLIP. For a given facial

image I , we extract its image feature I_{feat} using the image encoder $E_I(\cdot)$ of CLIP. Then, for each region-specific set D_{region} , we compute the cosine similarity between I_{feat} and every v_i in that set:

$$\text{sim}(I_{\text{feat}}, v_i) = \frac{I_{\text{feat}} \cdot v_i}{\|I_{\text{feat}}\| \|v_i\|} \quad (1)$$

We select the text feature with the highest similarity as the most relevant description for that region:

$$k_{\text{region}} = \arg \max_i \text{sim}(I_{\text{feat}}, v_i) \quad (2)$$

Finally, we concatenate the most relevant textual features across all regions to form the full semantic representation s of the face image, s serves as the region-wise aggregated semantic embedding for the image, capturing facial semantic attribute from different facial regions.

Adapter. We train a lightweight template→attribute alignment (TAA) adapter: $\mathcal{T} \rightarrow \mathcal{S}$ to approximate s from the facial template t , TAA consists of two fully-connected layers with hidden size h and a non-linearity ReLU. The loss combines element-wise MSE and cosine alignment:

$$\mathcal{L}_{\text{sem}} = \lambda_{\text{mse}} \|s - \hat{s}\|_2^2 + \lambda_{\text{cos}} (1 - \cos(s, \hat{s})),$$

with $\lambda_{\text{mse}} = 0.7$, $\lambda_{\text{cos}} = 0.3$. Optimization uses Adam (lr = 10^{-3}) for 20 epochs. After training, only TAA (not the image) is needed at attack time.

Fusion Mapping to Latent

After obtaining the facial template t and its predicted attribute embedding \hat{s} , we concatenate them with a noise vector $n \sim \mathcal{N}(0, I)$ and feed the fused representation $[n, t, \hat{s}]$ into the projector M_{FLP} , which maps it to a latent code $\hat{w} \in \mathcal{W}$. To make \hat{w} conform to the StyleGAN prior, M_{FLP} is jointly trained with a critic under a WGAN objective; the full training pipeline for the M_{FLP} and critic pair is illustrated in Figure. 3.

The architecture of M_{FLP} . We train the fusion mapping network M_{FLP} to project leaked facial template and semantic embedding plus controlled stochastic variation into the StyleGAN3 \mathcal{W} space. To effectively combine facial template and semantic embedding, we design M_{FLP} with a multi-branch feedforward architecture. Given a facial template t , its predicted semantic embedding \hat{s} which obtained by TAA, we first obtain projected representations t' and s' . The semantic embedding \hat{s} encodes facial feature attributes (eyes / nose / mouth / etc.); thus s' is partitioned into tokens $\{s'_r\}_{r=1}^R$. Since facial template t encodes highly discriminative identity information. We apply multi-head attention (MHA) with the template projection as query:

$$\tilde{s} = \text{MHA}(Q = t', K = [s'_1, \dots, s'_R], V = [s'_1, \dots, s'_R]),$$

producing an attention-refined semantic summary \tilde{s} emphasizing semantic embeddings most relevant to identity (e.g., eye shape, lip contour). The random vector n is normalized and directly used to provide stochastic variation. All three representations, the normalized noise vector, the projected template t' , and the attention-refined semantic summary \tilde{s} , are concatenated to form a unified representation

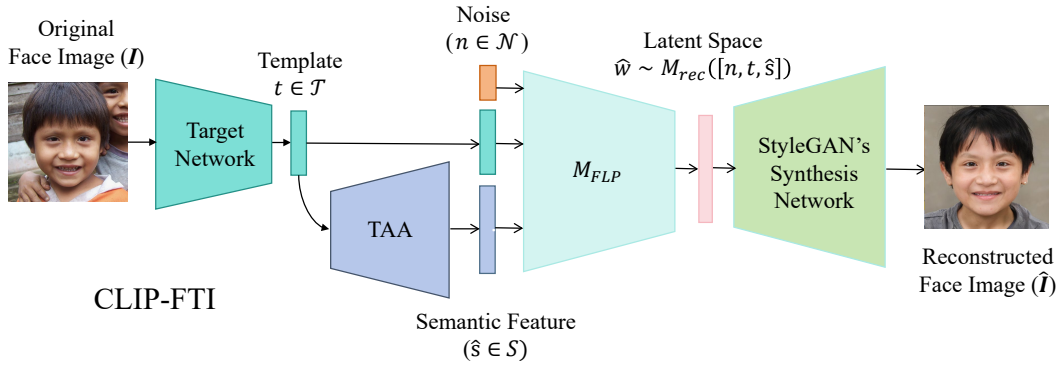


Figure 1: Overview of **CLIP-FTI**. An original image I yields a facial template t (ArcFace/ElasticFace). A template→semantic alignment adapter TAA learns to predict \hat{s} from t . A fusion mapping M_{FLP} takes noise n , t , and \hat{s} at attack time and outputs $\hat{w} \in \mathcal{W}$; the frozen StyleGAN3 synthesis network synthesizes a reconstructed image $\hat{I} = G(\hat{w})$.

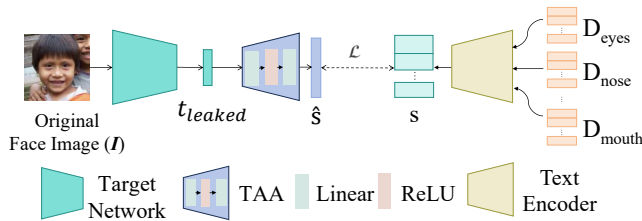


Figure 2: Training the template→attribute alignment adapter TAA . The adapter learns to predict the aggregated semantic embedding s from the facial template t using MSE and cosine loss.

z_f , which is then passed through a feedforward stack (MLP + LeakyReLU blocks) to obtain a StyleGAN latent:

$$\hat{w} = M_{FLP}(n, t, \hat{s}) \in \mathcal{W}.$$

Compared to previous methods mapping to \mathcal{W} space using t only, M_{FLP} enables the network to produce high-quality latent codes that reflect both semantic attributes and identity characteristics, fine-grained attribute conditioning via attention compensates for the expressivity otherwise gained from relaxing to \mathcal{W}^+ , narrowing the identity-detail trade-off. The architecture of the M_{FLP} is illustrated in Figure. 4.

Training Details

Latent distribution alignment. To keep generated latent codes on the StyleGAN prior manifold, we use a GAN-based framework based on Wasserstein GAN (WGAN) algorithm to learn the distribution of intermediate latent space \mathcal{W} of StyleGAN model. Real latent codes are sampled as $w_{\text{real}} = M_{\text{StyleGAN}}(z)$, $z \sim P_Z$. Generated latent codes are $\hat{w} = M_{FLP}(n, t, \hat{s})$. A critic $C : \mathcal{W} \rightarrow \mathbb{R}$ is trained with:

$$\mathcal{L}_C^{\text{WGAN}} = \mathbb{E}_{w_{\text{real}}} [C(w_{\text{real}})] - \mathbb{E}_{\hat{w}} [C(\hat{w})], \quad (3)$$

$$\mathcal{L}_{M_{FLP}}^{\text{WGAN}} = -\mathbb{E}_{\hat{w}} [C(\hat{w})]. \quad (4)$$

The critic C is a 3-layer MLP with LeakyReLU activations.

Reconstruction-guided refinement. Given the coarse latent \hat{w} predicted by M_{FTI} , the frozen generator produces a face

image $\hat{I} = G(\hat{w})$. We refine \hat{I} by minimising a composite loss that balances pixel fidelity, identity consistency, attribute agreement, and perceptual realism.

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{pix}} \mathcal{L}_{\text{pix}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{attr}} \mathcal{L}_{\text{attr}} + \lambda_{\text{perc}} \mathcal{L}_{\text{lpips}},$$

The final training objective therefore becomes

$$\mathcal{L}_{M_{FLP}}^{\text{total}} = \mathcal{L}_{M_{FLP}}^{\text{WGAN}} + \mathcal{L}_{\text{rec}}.$$

which jointly preserves latent realism and tightens the match between \hat{I} and the ground-truth facial attributes.

Implementation Details

We employ the Adam optimizer (Kingma and Ba 2017) with an initial learning rate of 1×10^{-1} and a `StepLR` scheduler that multiplies the learning rate by 0.5 every three epochs. The total loss is a weighted sum with coefficients $\lambda_{\text{pix}} = 1.0$, $\lambda_{\text{id}} = 1$, $\lambda_{\text{attr}} = 1$ and $\lambda_{\text{lpips}} = 1$. All experiments are conducted on a single NVIDIA RTX 3090 (24 GB). For the StyleGAN model, we use a pretrained model of StyleGAN3 in our experiments and generate 1024×1024 resolution face images.

Experiments

Datasets. Following a black-box threat model, our face reconstruction network CLIP-FTI is trained on FlickrFaces-HQ (FFHQ) (Karras, Laine, and Aila 2019) dataset, which contains 70,000 high-quality unlabeled images. We randomly split FFHQ into 90% training and 10% test sets. For evaluation, we use three datasets: LFW (Karras, Laine, and Aila 2019), CelebA-HQ (Karras et al. 2017), and AgeDB (Moschoglou et al. 2017). LFW contains 13,233 images of 5,749 subjects; CelebA-HQ includes 30,000 aligned celebrity faces at 1024×1024 resolution; and AgeDB offers 16,488 images of 568 individuals with large age variations (up to ~ 50 years per identity).

Models. In our experiments, we consider different SOTA FR models including ArcFace (Deng et al. 2019), ElasticFace (Boutros et al. 2022) as well as three different FR models with SOTA backbones from FaceX-Zoo (Wang et al.

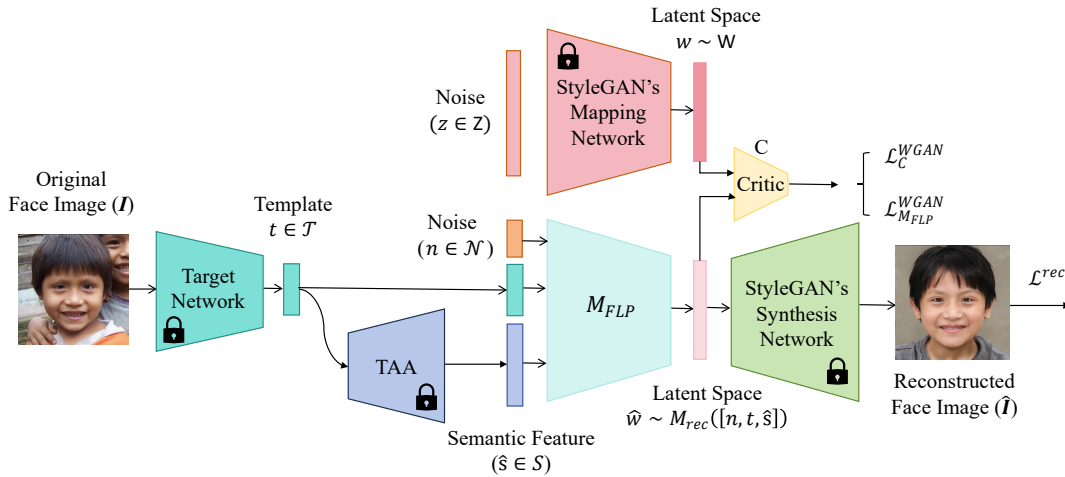


Figure 3: Training of the fusion-latent projector M_{FLP} and critic. A leaked template t , its TAA-predicted attribute embedding \hat{s} , and noise n are fused by M_{FLP} to produce a latent \hat{w} ; the critic aligns \hat{w} with the StyleGAN latent prior, while reconstruction losses between the generated image $\hat{I} = G(\hat{w})$ and the original image I further supervise M_{FLP} .

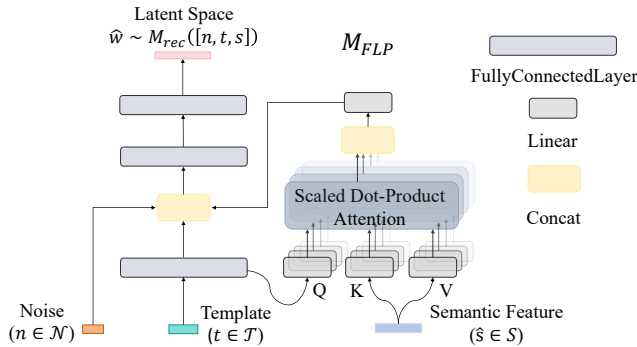


Figure 4: Internal architecture of M_{FLP} : three branches (noise n , template t , semantic embeddings \hat{s}) \rightarrow attention fusion \rightarrow MLP \rightarrow latent $w \in W$.

2021), including HRNet (Wang et al. 2020), AttentionNet (Wang et al. 2017), and GhostNet (Han et al. 2020). All FR models are trained on MS-Celeb1M (Guo et al. 2016).

Metrics. For evaluation with each of the LFW, CelebA-HQ and AgeDB datasets, we build a separate FR system, and use the reference templates (i.e., enrolled in the system’s database) as input to our face reconstruction method. Then we inject the reconstructed face image as a query to the system and evaluate four complementary criteria:

(i) Verification Protocol. We use BLUFR (Liao et al. 2014) to verify the accuracy of identification. A gallery \mathcal{G} of enrolled templates is fixed. Cosine similarity between ℓ_2 -normalized embeddings provides the match score. The decision threshold is tuned on impostors to meet $FAR = \{0.01\%, 0.1\%, 1\%\}$; we then report TAR at the same points.

- **Type-I** (P_I): reconstructions of the *enrolled* images.
- **Type-II** (P_{II}): reconstructions from a different image of the same identity.
- **Impostor** (P_N): images whose identities are absent from

\mathcal{G} (used only for threshold setting).

We extend BLUFR to CelebA-HQ and AgeDB by forming cross-identity (impostor) pairs via the Cartesian product, ensuring robust impostor statics and BLUFR-style evaluation. Then we report Type-I and Type-II TAR on each dataset and model at three FAR levels. The details are provided in the supplementary materials.

(ii) MS-SSIM. Multiscale Structural Similarity (MS-SSIM) (Wang, Simoncelli, and Bovik 2003) evaluates luminance, contrast, and structural agreement across a cascade of down-sampled resolutions, yielding a holistic perceptual score that correlates well with human judgments on both low and high frequency content. Because it penalises structural distortions at every scale, MS-SSIM is particularly suited for assessing the global realism of GAN-based reconstructions. Values lie in $[0, 1]$; larger is better.

(iii) LPIPS. Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) measures the ℓ_2 distance between deep features extracted from a fixed, pretrained network, producing a scalar in $[0, 1]$ where smaller is better.

(iv) FAMSE. Facial Attribute Mean-Squared Error (FAMSE) quantifies fine-grained semantic fidelity. For each image pair (x, x') we first localise eyes, nose, and mouth via 68-point landmarks (Zhang et al. 2016); a lightweight attribute encoder then predicts a 512-D semantic vector for every region. The MSE between corresponding vectors is averaged over the three regions to obtain FAMSE, with lower scores signalling closer alignment in component-level attributes and, by extension, stronger identity consistency.

Experimental Results

Identification accuracy and attribute similarity. Table 2 demonstrates that *CLIP-FTI* consistently outperforms the strongest published inversion baseline on all three bench-

Method	$F_{\text{database}} / F_{\text{loss}}$	LFW				CelebA-HQ		AgeDB	
		TAR@FAR=0.1%		TAR@FAR=1%		TAR@FAR=0.1%	TAR@FAR=1%	TAR@FAR=0.1%	TAR@FAR=1%
		Type-I \uparrow	Type-II \uparrow	Type-I \uparrow	Type-II \uparrow	Type-I \uparrow	Type-I \uparrow	Type-I \uparrow	Type-I \uparrow
Otroshi <i>et al.</i>	ArcFace/	0.9501	0.4655	0.9933	0.8125	0.8979	0.9827	0.7982	0.9405
Ours	ElasticFace	0.9937	0.8174	0.9995	0.9548	0.9535	0.9937	0.9002	0.9787
Otroshi <i>et al.</i>	ElasticFace/	0.9486	0.5861	0.9893	0.8336	0.9691	0.9960	0.9226	0.9743
Ours	ArcFace	0.9435	0.5956	0.9889	0.8517	0.9715	0.9967	0.9503	0.9898

Table 2: Type-I/II TAR (%) on LFW and Type-I TAR on CelebA-HQ and AgeDB at two FAR operating points (0.1% and 1%). F_{database} denotes the model that generated the leaked template, and F_{loss} is the surrogate used during training. For CelebA-HQ and AgeDB, we only report Type-I results.

Method	LFW				CelebA-HQ				AgeDB			
	MS-SSIM \uparrow	LPIPS \downarrow	MSE \downarrow	FAMSE \downarrow	MS-SSIM \uparrow	LPIPS \downarrow	MSE \downarrow	FAMSE \downarrow	MS-SSIM \uparrow	LPIPS \downarrow	MSE \downarrow	FAMSE \downarrow
Otroshi <i>et al.</i>	0.2428	0.5534	0.0766	0.0503	0.1927	0.5762	0.1014	0.1001	0.2081	0.6064	0.1019	0.0473
Ours	0.2527	0.5419	0.0662	0.0451	0.2148	0.5571	0.0880	0.0975	0.2377	0.5870	0.0884	0.0437

Table 3: Quantitative comparison across datasets. \uparrow indicates higher is better, \downarrow lower is better.

marks and at both FAR operating points. In the canonical ArcFace \rightarrow ElasticFace setting, our method markedly elevates not only the conventional Type-I verification rate but also the more demanding Type-II metric, and these gains remain stable after the database and surrogate networks are swapped. The fact that the same pattern recurs on CelebA-HQ and AgeDB indicates that the CLIP-conditioned latent search is model-agnostic and does not overfit to any single data distribution. Further results details are included in the supplementary material.

The qualitative evidence in Figure 5 echoes these quantitative results. Whereas the baseline occasionally drifts in high-level semantics (e.g., gender or overall facial attributes) *CLIP-FTI* produces reconstructions that retain both global identity and fine-grained attributes across all datasets. Together, the metrics and visuals confirm that our approach obtains state-of-the-art identification accuracy while faithfully preserving the semantic attributes of reconstructed faces.

Component-level attribute semantics. As summarized in Table 3, *CLIP-FTI* delivers consistent improvements over the previous best attack across all three benchmarks and every metric considered. Most notably, it reduces Facial Attribute MSE (FAMSE) (our most discriminative measure of eyes, nose, and mouth fidelity) on each dataset, indicating more accurate recovery of component-level semantics. These gains come hand-in-hand with higher MS-SSIM and lower perceptual distance scores, confirming that the method preserves or enhances global structure and visual realism while sharpening semantic details. Complementary reductions in pixel domain error further underscore the holistic quality of the reconstructions. The consistently positive margins observed across all benchmarks underscore the robustness of the CLIP-guided latent search. Collectively, the results demonstrate that injecting textual attribute embeddings into the inversion pipeline yields faces that are both perceptually faithful and semantically precise, thereby strengthen-

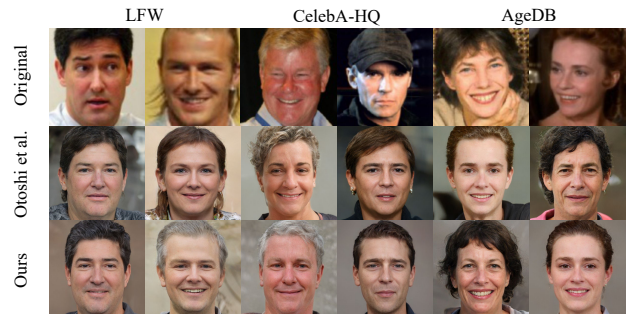


Figure 5: Qualitative comparison across datasets and reconstruction methods. The first row shows the original images, the second row shows reconstructions by *Otroshi et al.*, and the third row presents the results of our CLIP-FTI. Column groups (left to right) correspond to two examples each from LFW, CelebA-HQ, and AgeDB.

ing the practical effectiveness of template-only attacks and amplifying the attendant privacy risks.

Cross-model attack transferability. Table 4 summarizes 30 transfer scenarios obtained by crossing two leak surrogate pairs (ArcFace \rightarrow ElasticFace and ElasticFace \rightarrow ArcFace) with five target FR and three datasets. *CLIP-FTI* outperforms the strongest prior attack in all but two cases; in those exceptions (the ArcFace and ElasticFace self-tests under the ElasticFace \rightarrow ArcFace regime) our method performs virtually the same as the baseline, remaining within 0.6 percentage points. On average, it delivers double digit relative gains on the most challenging cross-architecture pairs while never regressing on any benchmark. Crucially, the largest gaps appear on the structurally dissimilar lightweight backbones HRNet, AttentionNet, and GhostNet, underscoring that the CLIP-

F_{database}	F_{loss}	F_{target}	LFW		CelebA-HQ		Age-DB	
			Otroshi <i>et al.</i>	Ours	Otroshi <i>et al.</i>	Ours	Otroshi <i>et al.</i>	Ours
ArcFace	ElasticFace	ArcFace	95.10	99.37	89.79	95.35	79.82	94.05
		ElasticFace	83.90	91.62	85.31	91.84	58.04	68.37
		HRNet	51.63	65.23	44.29	59.15	6.00	9.80
		AttentionNet	33.99	43.56	14.26	23.06	0.96	2.04
		GhostNet	18.72	27.58	15.22	22.97	0.97	1.30
ElasticFace	ArcFace	ArcFace	96.41	96.20	95.22	95.43	67.16	69.25
		ElasticFace	94.86	94.35	96.91	97.15	99.60	99.67
		HRNet	77.35	77.88	62.99	64.81	15.66	18.57
		AttentionNet	52.18	52.54	32.02	32.67	3.53	4.37
		GhostNet	38.58	38.92	24.00	28.87	3.03	3.72

Table 4: Cross-model transferability. Type-I TAR (%) at FAR = 10^{-3} on three benchmarks for every combination of leaked backbone (F_{database}), surrogate backbone (F_{loss}), and target backbone (F_{target}). We compare the strongest prior attack (Otroshi *et al.*) with our CLIP-FTI.

Variant	Type-I \uparrow	Type-II \uparrow	MS-SSIM \uparrow	MSE \downarrow	FAMSE \downarrow	LPIPS \downarrow
CLIP-FTI	0.9937	0.8174	0.2527	0.0662	0.0451	0.5419
w/o AttrEmb	0.9510	0.4655	0.2428	0.0766	0.0503	0.5534
w/o MHA	0.9553	0.4712	0.2443	0.0756	0.0501	0.5520
w/o ConMHA	0.9619	0.4792	0.2483	0.0742	0.0498	0.5501

Table 5: Ablation on architectural components (LFW, FAR= 10^{-3}).

Variant	Type-I \uparrow	Type-II \uparrow	MS-SSIM \uparrow	MSE \downarrow	FAMSE \downarrow	LPIPS \downarrow
CLIP-FTI	0.9822	0.7229	0.2527	0.0662	0.0451	0.5419
w/o $\mathcal{L}_{\text{attr}}$	0.9819	0.6447	0.2563	0.0654	0.0428	0.5393
w/o $\mathcal{L}_{\text{lips}}$	0.9618	0.4453	0.2248	0.0753	0.0450	0.5593
w/o both	0.9783	0.5085	0.2261	0.0829	0.0495	0.5478

Table 6: Ablation on loss terms (LFW, FAR= 10^{-3}).

conditioned latent search does not hinge on architectural similarity between the surrogate and the target. In summary, a single *CLIP-FTI* model can successfully attack face recognition systems that differ widely in scale, training loss, and architecture, broadening the real world threat and underscoring the need for backbone-agnostic defense strategies.

Ablation Study

Effect of CLIP-driven attribute embeddings and cross-modal attention. Table 5 yields three observations. (i) *Embeddings matter.* Removing CLIP-driven attribute embeddings (w/o AttrEmb) causes the largest drop across all metrics, showing that a template-only signal lacks the component-level semantics needed for cross-view verification. (ii) *Attention matters.* Simple concatenation (w/o MHA) recovers little, indicating the need for an explicit mechanism to adaptively weight attribute cues. (iii) *Query choice matters most.* Using MHA without the template as query (w/o ConMHA) improves over concatenation but trails the full model; making the template guide the attention is crucial for focusing on identity-determining parts (eyes, nose, mouth) and fully exploiting the attribute embeddings.

Effect of attribute and perceptual losses. Table 6 indicates that the two loss terms address complementary aspects of reconstruction. Without either loss, the generator converges to over-smoothed images and performs poorly on all metrics. Adding only the *attribute loss* (w/o $\mathcal{L}_{\text{lips}}$) emphasizes global traits from an identity-agnostic extractor (e.g., age, expression), lowering pixel-space errors (MSE, FAMSE) but yielding little change in Type-I/II TAR. Applying only the *perceptual loss* (w/o $\mathcal{L}_{\text{attr}}$) improves local texture/structure and verification accuracy, while fine attributes drift and pixel metrics remain largely unchanged. Optimizing both losses balances these effects: the attribute loss provides semantic guidance, and the perceptual loss restores high-frequency detail, yielding the best verification accuracy with a minor increase in pixel error—an acceptable trade-off when the goal is to pass FR checks.

Effect of CLIP attribute guidance on local regions. To verify that our CLIP attribute embeddings act on the intended facial parts, we compute region-wise CLIP cosine similarity between (i) images produced by Otroshi *et al.* and (ii) images produced by ours, and the corresponding attribute embeddings (eyes, nose, mouth, jaw, eyebrow). The detail results are provided in the supplementary materials.

Conclusion

In this paper, we revisited face template inversion (FTI) through fine-grained attribute recovery and proposed **CLIP-FTI**. By predicting component-level facial attribute embeddings from a leaked template and fusing them with the template via a cross-modal attention network, and project to StyleGAN \mathcal{W} space to synthesize identity-consistent, attribute-faithful faces. Extensive experiments on multiple datasets and FR backbones show that **CLIP-FTI** achieves a new state-of-the-art across all attack metrics on standard benchmarks while retaining strong cross-model transferability. To the best of our knowledge, we are the first to introduce external textual semantic information into the face template inversion attack task, enabling more precise and controllable attribute reconstruction across diverse facial regions.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grant numbers, U23B2023 and 62472199, Guangdong Key Laboratory of Data Security and Privacy Preserving under Grant 2023B1212060036, the basic and Applied Basic Research Foundation of Guangdong Province (2025A1515011097), and the Outstanding Youth Project of Guangdong Basic and Applied Basic Research Foundation (2023B1515020064). This work is also supported by Engineering Research Center of Trustworthy AI, Ministry of Education.

References

- Abdal, R.; Zhu, P.; Femiani, J.; Mitra, N. J.; and Wonka, P. 2021. CLIP2StyleGAN: Unsupervised Extraction of StyleGAN Edit Directions. *CoRR*, abs/2112.05219.
- Biggio, B.; Russu, P.; Didaci, L.; Roli, F.; et al. 2015. Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective. *IEEE Signal Processing Magazine*, 32(5): 31–41.
- Boutros, F.; Damer, N.; Kirchbuchner, F.; and Kuijper, A. 2022. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1578–1587.
- Cole, F.; Belanger, D.; Krishnan, D.; Sarna, A.; Mosseri, I.; and Freeman, W. T. 2017. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3703–3712.
- Deb, D.; Zhang, J.; and Jain, A. K. 2020. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. IEEE.
- 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.
- Dong, X.; Jin, Z.; Guo, Z.; and Teoh, A. B. J. 2021. Towards generating high definition face images from deep templates. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1–11. IEEE.
- Dong, X.; Miao, Z.; Ma, L.; Shen, J.; Jin, Z.; Guo, Z.; and Teoh, A. B. J. 2023. Reconstruct face from features based on genetic algorithm using GAN generator as a distribution constraint. *Computers & Security*, 125: 103026.
- Fei, J.; Dai, Y.; Yu, P.; Shen, T.; Xia, Z.; and Weng, J. 2022. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20270–20280.
- Fei, J.; Xia, Z.; Yu, P.; and Xiao, F. 2020. Adversarial attacks on fingerprint liveness detection. *EURASIP Journal on Image and Video Processing*, 2020(1): 1.
- 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.
- Galbally, J.; Marcel, S.; and Fierrez, J. 2014. Biometric anti-spoofing methods: A survey in face recognition. *Ieee Access*, 2: 1530–1552.
- Galbally, J.; McCool, C.; Fierrez, J.; Marcel, S.; and Ortega-Garcia, J. 2010. On the vulnerability of face verification systems to hill-climbing attacks. *Pattern Recognition*, 43(3): 1027–1038.
- Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, 87–102. Springer.
- Hadid, A.; Evans, N.; Marcel, S.; and Fierrez, J. 2015. Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. *IEEE Signal Processing Magazine*, 32(5): 20–30.
- Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; and Xu, C. 2020. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1580–1589.
- Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.
- 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.
- Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.
- Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*.
- Liao, S.; Lei, Z.; Yi, D.; and Li, S. Z. 2014. A benchmark study of large-scale unconstrained face recognition. In *IEEE international joint conference on biometrics*, 1–8. IEEE.
- Mai, G.; Cao, K.; Yuen, P. C.; and Jain, A. K. 2018. On the reconstruction of face images from deep face templates. *IEEE transactions on pattern analysis and machine intelligence*, 41(5): 1188–1202.
- Marcel, S.; Fierrez, J.; and Evans, N. 2023. *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, volume 1. Springer.
- Marcel, S.; Nixon, M. S.; Fierrez, J.; and Evans, N. 2019. *Handbook of biometric anti-spoofing: Presentation attack detection*, volume 2. Springer.
- Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; and Zafeiriou, S. 2017. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 51–59.
- Otroshi Shahreza, H.; and Marcel, S. 2023. Face reconstruction from facial templates by learning latent space of a generator network. *Advances in Neural Information Processing Systems*, 36: 12703–12720.

Paraperas Papantoniou, F.; Lattas, A.; Moschoglou, S.; Deng, J.; Kainz, B.; and Zafeiriou, S. 2024. Arc2Face: A Foundation Model for ID-Consistent Human Faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

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.

Shahreza, H. O.; and Marcel, S. 2023. Comprehensive vulnerability evaluation of face recognition systems to template inversion attacks via 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14248–14265.

Shahreza, H. O.; and Marcel, S. 2024. Template inversion attack using synthetic face images against real face recognition systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Vendrow, E.; and Vendrow, J. 2021. Realistic face reconstruction from deep embeddings. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, J.; Liu, Y.; Hu, Y.; Shi, H.; and Mei, T. 2021. Facex-zoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM international conference on multimedia*, 3779–3782.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multi-scale structural similarity for image quality assessment. In *The thirty-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, 1398–1402. Ieee.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.