

TriAnchor-ID: Multi-View Semantic Identity Anchoring for Pose-Aware Personalized Diffusion

A Prototype Study on Identity Consistency in Diffusion Portrait Generation

Akash Kumar

April 2026

Abstract

Personalized diffusion models often suffer from identity drift: generated portraits may remain recognizable under simple prompts but change noticeably under pose, expression, lighting, or style variation. This paper presents TriAnchor-ID, a prototype framework for identity-conditioned portrait generation based on multi-view face identity extraction, quality-weighted embedding aggregation, and explicit post-generation identity evaluation. The central correction is mathematical and methodological: the implemented 512-dimensional ArcFace vector is treated as a semantic identity anchor, denoted e_{id} , rather than as a physical 3DMM/FLAME shape parameter. The proposed Identity Capsule separates implemented semantic identity evidence from geometric observations and from future extensions such as FLAME shape fitting, UV-space texture, and spatial control branches. The accompanying prototype implements multi-view identity extraction, landmark-based geometric analysis, capsule serialization, IP-Adapter FaceID generation hooks, and an ArcFace-based evaluation scaffold. The work is positioned as an evidence-aligned experimental prototype, not as a claim of guaranteed biometric consistency or state-of-the-art performance.

Keywords: personalized diffusion, identity drift, ArcFace, InsightFace, IP-Adapter FaceID, SDXL, ControlNet, FLAME, identity-conditioned generation

1. Introduction

Text-to-image diffusion systems can synthesize realistic portraits, yet preserving the same person across changing prompts remains difficult. A model may preserve a loose resemblance under frontal, well-lit prompts but drift when asked for profile views, unusual lighting, stylized output, or strong scene changes. This problem is referred to here as identity drift.

TriAnchor-ID addresses this problem as a prototype identity-conditioning framework. Instead of describing facial identity as a single physical vector, the framework separates identity evidence into semantic recognition, landmark geometry, local facial regions, texture information, and confidence metadata. The present implementation focuses on the semantic anchor and landmark geometry. Future versions can extend the capsule with fitted 3DMM/FLAME shape and UV texture.

The strongest claim supported by the current implementation is that multi-view semantic identity extraction and quality-weighted embedding averaging provide a measurable foundation for identity-conditioned generation. Quantitative claims about superiority require controlled repeated experiments against baselines.

1.1 Contributions

- A corrected formulation that distinguishes the implemented ArcFace semantic identity anchor e_{id} from future physical 3DMM/FLAME shape parameters.
- An Identity Capsule design that organizes semantic identity, landmark geometry, quality weights, metadata, and future geometry/texture extensions.

- A multi-view aggregation method that averages normalized identity embeddings using quality-aware weights rather than relying on a single reference image.
- A practical evaluation protocol using ArcFace cosine similarity, identity drift, landmark error, detection failure rate, and prompt-alignment metrics.
- A conservative scope of claims that avoids unsupported guarantees and treats qualitative images as illustrative until evaluated under repeated seed-controlled conditions.

2. Related Work

The prototype builds on several established directions. Latent diffusion models provide the base text-to-image generation mechanism [1], while SDXL improves large-scale text-to-image synthesis quality [2]. IP-Adapter introduces a lightweight image-prompt adapter compatible with text prompts [3], and IP-Adapter FaceID variants use face identity embeddings for identity-conditioned generation. ArcFace provides an additive angular-margin training objective for face recognition embeddings [4]. ControlNet demonstrates how spatial conditions can be injected into diffusion models through trainable control branches initialized to preserve the base model behavior [5]. FLAME provides a parametric 3D model of facial shape and expression, which is useful as a future extension but is not yet fitted by the present notebook [6].

3. Method

3.1 Corrected Prototype Pipeline

The implemented pipeline extracts identity and landmark evidence from multiple reference views, aggregates the semantic identity embeddings, stores a structured identity capsule, and uses the resulting identity anchor for identity-conditioned generation.

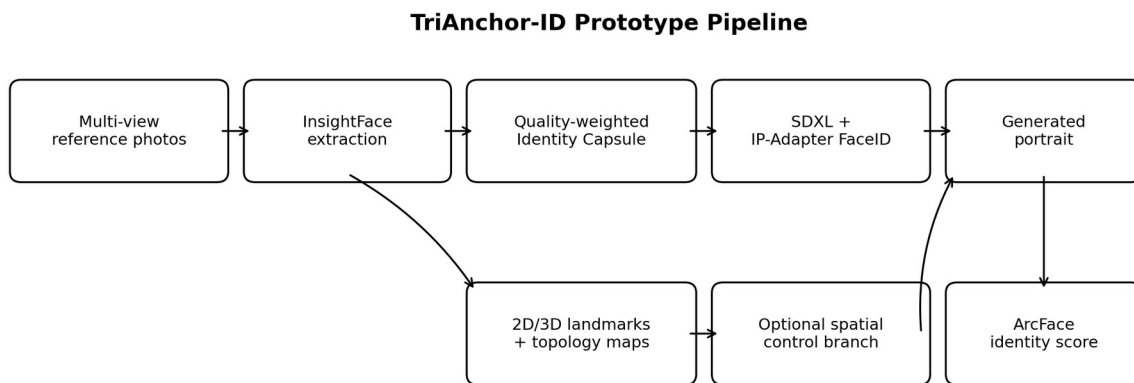


Figure 1. TriAnchor-ID prototype pipeline. The implemented path uses multi-view identity extraction, quality-weighted identity aggregation, IP-Adapter FaceID generation hooks, and ArcFace-based post-generation evaluation. The spatial control branch is a planned extension.

Stage	Output	Scientific interpretation
Reference upload	3-8 face images	Multi-view evidence for the same subject.
InsightFace extraction	ArcFace embedding, bounding box, 2D/3D landmarks	Semantic identity evidence plus geometric observations.

Stage	Output	Scientific interpretation
Capsule construction	e_id, landmarks, bounding boxes, quality weights, metadata	Implemented Identity Capsule prototype.
Generation	SDXL + IP-Adapter FaceID hooks	Semantic identity-conditioned generation.
Evaluation	ArcFace cosine, drift score, detection status	Measurable identity consistency after generation.

3.2 Identity Capsule

Let C_ID denote the Identity Capsule. The capsule is defined as a product-space object rather than a single loosely named vector:

```
C_ID in A x T x E x L x M

alpha_shape in R^{d_s}           # future FLAME/3DMM shape parameter
tau_texture in R^{H_uv x W_uv x C} # future UV-space texture representation
e_id in S^{d_e - 1}              # implemented ArcFace semantic embedding
L_local = {l_r in S^{d_r - 1}}_{r=1}^R # future or derived local region descriptors
m_conf in [0,1]^{H_uv x W_uv}    # confidence / visibility mask
```

In the current prototype, e_id is implemented and serialized as a normalized 512-dimensional semantic identity anchor. Landmark-derived geometric evidence is also saved in the capsule file. However, $alpha_shape$ and $tau_texture$ are not fitted in the present implementation. A full geometry-aware version should add FLAME, DECA, or EMOCA fitting for $alpha_shape$ and UV unwrapping or neural texture extraction for $tau_texture$.

3.3 Quality-Weighted Multi-View Identity Aggregation

For each reference image x_i , the face recognition model extracts a normalized embedding e_i and quality metadata q_i . Quality can include detector confidence, face size, pose visibility, blur estimates, and landmark reliability. The semantic identity anchor is computed as:

```
e_i = F_ID(x_i) / ||F_ID(x_i)||_2
w_i = q_i / sum_j q_j
e_id = normalize(sum_i w_i * e_i)
```

The purpose of multi-view aggregation is not to create a physical face model. It is to reduce dependence on a single reference view and provide a more stable semantic identity anchor for downstream conditioning and evaluation.

4. Evaluation Metrics

Generated images are denoted $\hat{x} = G_theta(z, p, C_ID)$, where z is the random seed/noise, p is the text prompt, and C_ID is the identity capsule.

```
Identity drift:
D_ID(x_hat, C_ID) = 1 - cos(F_ID(x_hat), e_id)

Expected identity drift over prompts and seeds:
D_ID_expected(G_theta, C_ID) = E_{p ~ P, z ~ N(0, I)} [1 - cos(F_ID(G_theta(z, p, C_ID)), e_id)]

Geometry error using normalized landmark distance:
```

```

D_geo(x_hat, C_ID) = (1 / (K * d_face^2)) * sum_i || l_i(x_hat) - l_i(C_ID) ||_2^2

Local-region drift:
D_local(x_hat, C_ID) = sum_r w_r * [1 - cos(F_r(x_hat), l_r)]

Prompt-alignment error:
D_prompt(x_hat, p) = 1 - CLIP(x_hat, p)

```

These metrics deliberately separate identity preservation from prompt adherence. A model that preserves identity while ignoring the prompt is not successful; neither is a model that follows the prompt while losing the subject identity.

5. Theoretical Motivation: Conditional Drift Bound

The following bound is a theoretical motivation for capsule quality, not a proof of perfect identity preservation. It depends on assumptions that must be validated or approximated empirically.

Assume C_ID^* is an ideal identity capsule and C_ID is the extracted capsule. Suppose G_theta is locally Lipschitz with respect to capsule perturbations, and F_ID is L_F -Lipschitz with respect to image perturbations. If extraction errors are bounded by ϵ_{shape} , $\epsilon_{texture}$, $\epsilon_{semantic}$, and ϵ_{local} , then the identity drift caused by capsule error can be upper-bounded by a weighted sum of those errors.

```

Assumptions:
||delta_alpha|| <= epsilon_shape
||delta_tau|| <= epsilon_texture
||delta_e|| <= epsilon_semantic
sum_r ||delta_l_r|| <= epsilon_local

Conditional drift bound:
D_ID(G_theta(z, p, C_ID), G_theta(z, p, C_ID*)) <=
  K_shape * epsilon_shape
+ K_texture * epsilon_texture
+ K_semantic * epsilon_semantic
+ K_local * epsilon_local

```

Proof sketch. By local Lipschitz continuity of the generator, the image-space change is bounded by a weighted sum of capsule-component perturbations. Applying the Lipschitz continuity of the identity encoder maps image perturbation into embedding perturbation. For normalized embeddings, cosine distance can be bounded by embedding displacement up to a constant factor. This argument supports the design principle that lower extraction error should reduce drift under controlled conditions; it does not guarantee identity consistency.

6. Implementation Status

Component	Implemented?	Role
ArcFace e_id	Yes	Global semantic identity anchor for IP-Adapter FaceID.
2D/3D landmarks	Yes	Geometric evidence and topology visualization.
Quality weighting	Yes	Downweights weak reference images during embedding aggregation.

Component	Implemented?	Role
Capsule serialization	Yes	Stores identity anchor, landmarks, boxes, weights, and metadata.
FLAME alpha_shape	No - planned extension	Parametric 3D shape identity anchor.
UV tau_texture	No - planned extension	Texture and high-frequency identity details.
ControlNet spatial branch	No - planned extension	Depth, normal, landmark, and silhouette adherence.
Face consistency critic	No - planned extension	Inference-time identity drift correction.

7. Pilot Implementation Evidence

The corrected prototype initializes InsightFace, loads face-recognition models, extracts embeddings and landmarks from five reference views, creates a quality-weighted averaged semantic anchor `e_id`, and writes a structured identity capsule. The reported `e_id` shape is (512,), normalized. This is a valid semantic identity anchor and should not be described as a raw geometric alpha vector.

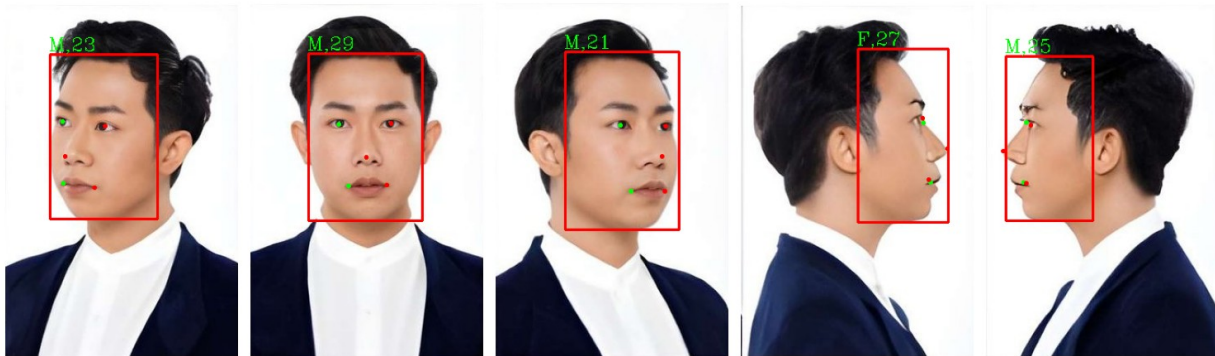


Figure 2. Multi-view reference detections. Bounding boxes and landmark points were detected across frontal, three-quarter, and profile views. These detections provide semantic and landmark evidence for the same subject.

Detailed Front-Facing Facial Topology (V)
(Geometric Skeleton & Surface)

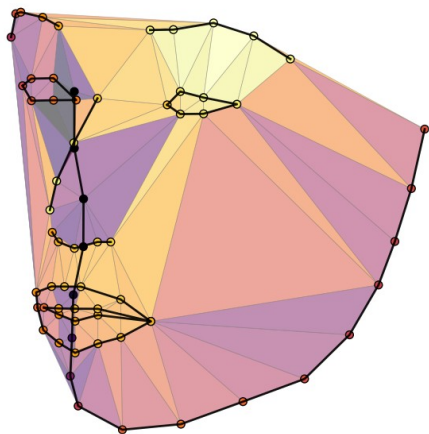


Figure 3. Landmark-derived facial topology. This is a landmark/topology visualization, not a fitted FLAME `alpha_shape`.

Qualitative generated portraits can be useful for human inspection, but they are not sufficient to support quantitative claims. A valid benchmark must run identical prompts, seeds, reference sets, and model settings across baselines, then report numeric metrics with confidence intervals.

8. Experimental Protocol for Credible Evaluation

A controlled benchmark should compare the proposed method against baselines across identical prompts, seeds, reference images, and pose/style conditions. Each reported chart should include sample size, mean, standard deviation, and 95% confidence intervals.

Item	Recommended requirement
Identities	Prototype: 1-5 identities. Stronger report: at least 50-100 identities.
References per identity	3-8 images, including frontal, three-quarter, and profile views.
Prompts per identity	10-20 prompts covering pose, lighting, expression, and style variation.
Seeds per prompt	At least 3-5 seeds per prompt.
Methods	Text-only, single-reference FaceID, multi-view e_id, and multi-view e_id plus spatial control.
Metrics	ArcFace similarity, identity drift, landmark NME, CLIP score, face-detection failure rate, and runtime.

Item	Recommended requirement
Charts	Use error bars or confidence intervals. Avoid unsupported recognition-threshold lines unless calibrated on the exact model and dataset.

9. Scope of Claims

Submission-safe wording	Rationale
Reduces identity drift under measured conditions.	Avoids unsupported guarantees such as sub-pixel identity consistency.
e_id is an ArcFace semantic embedding; alpha_shape is a future 3DMM parameter.	Correctly separates semantic identity from physical geometry.
Biases generation toward identity-consistent outputs.	Avoids claiming strict geometric enforcement before spatial branch training.
Requires controlled benchmark comparison before state-of-the-art claims.	Prevents overclaiming from qualitative examples alone.

10. Ethics and Security

Identity Capsules contain sensitive face-identity information. Even a semantic embedding such as e_id should be treated as private data. A product-grade capsule format should include encryption, subject consent metadata, audit logs, watermarking, and revocation mechanisms. The system should not be used to generate public figures or private individuals without consent. Evaluation datasets should be licensed and documented, and generated outputs should be labeled when appropriate.

11. Limitations

- The current prototype does not fit a true FLAME, BFM, DECA, or EMOCA alpha_shape.
- The current prototype does not extract a UV neural texture tau_texture.
- The current prototype does not train a ControlNet spatial branch.
- The current prototype does not implement inference-time identity critic guidance.
- Qualitative generated images alone are not enough for quantitative conclusions.
- ArcFace thresholds must be calibrated on the exact model, dataset, and deployment condition before being interpreted as recognition thresholds.

12. Conclusion

TriAnchor-ID presents a conservative and implementable prototype for identity-conditioned diffusion portrait generation. The work corrects the key mathematical distinction between a semantic ArcFace identity anchor and a physical 3DMM shape parameter, defines an extensible Identity Capsule, and provides a practical path for measuring identity drift. The present implementation supports multi-view semantic identity anchoring and landmark-based geometric analysis. The decisive next step is a controlled quantitative benchmark across prompts, poses, seeds, baselines, and identities, followed by spatial-control integration for geometry-aware conditioning.

References

- [1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR, 2022.
- [2] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Muller, J., Penna, J., and Rombach, R. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv, 2023.
- [3] Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv, 2023.
- [4] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. CVPR, 2019.
- [5] Zhang, L., Rao, A., and Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. ICCV, 2023.
- [6] Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. Learning a Model of Facial Shape and Expression from 4D Scans. ACM Transactions on Graphics / SIGGRAPH Asia, 2017.