SapiensID 2.0: Aligning Human Recognition Foundation Models with Human Perception

Yiyang Su[1] **Jie Zhu**[1] **Feng Liu**[2] **Anil Jain**[1] **Xiaoming Liu**[1,3]
[1] Michigan State University [2] Drexel University [3] University of North Carolina, Chapel Hill
{suyiyan1, zhujie4, jain}@msu.edu, fl397@drexel.edu, liuxm@cs.unc.edu

Abstract

While foundation models have significantly advanced human recognition across diverse modalities, they predominantly rely on static, geometric feature extraction. This approach fundamentally diverges from human perception. Consequently, current models often suffer from "semantic blindness," overfitting to transient noise while failing to leverage invariant soft biometrics, and struggle to capture temporal motion signatures. To bridge this gap, we propose SapiensID 2.0, a human recognition framework enriched with both semantic and temporal awareness. To overcome the lack of soft-biometric annotations, we transfer zero-shot semantic knowledge from Multimodal Large Language Models (MLLMs) into a discriminative embedding space. We resolve the dimensional mismatch between these spaces using Invariant Trait Alignment (ITA) to distill core persistent traits, and Transient Noise Disentanglement (TND) to decouple artifacts like clothing. Furthermore, we design a Kinematic Semantic Attention Head (K-SAH) that extends spatial attention across temporal windows. By tracking semantic patches over time, K-SAH captures rich kinematic signatures without requiring large-scale video datasets. Extensive experiments demonstrate that SapiensID 2.0 achieves state-of-the-art performance across image- and video-based person re-identification and gait recognition, while maintaining robust face recognition capabilities.

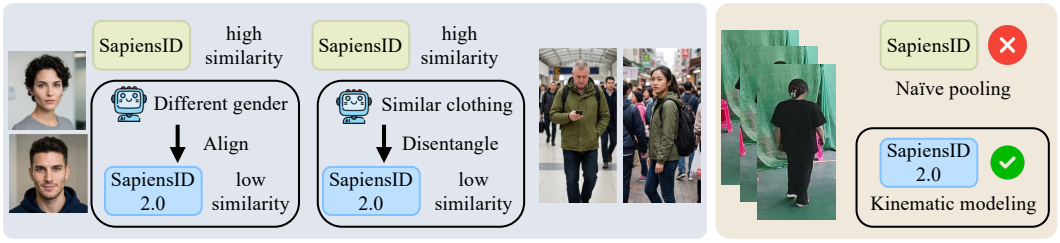
1 Introduction

Human recognition systems [42] typically isolate different modalities into distinct, modality-specific models. SapiensID [32] marked a paradigm shift by using a single model for face recognition and person re-identification (reID). In real-world scenarios, SapiensID can dynamically fall back on the other modality when extreme poses, poor visibility, or occlusions compromise a single modality (*e.g.*, an obscured face). However, SapiensID extracts features from raw visual textures of individual frames, so its embedding space lacks explicit semantic awareness and kinematic continuity—the two pillars of human recognition.

First, humans rely on interpretable semantic attributes [20]. We anchor identity to *fixed soft biometrics*—anatomically invariant traits such as race, gender, and body proportions. Also, we actively discount *dynamic soft biometrics*—transient, variable traits such as clothing or accessories. As shown in Fig. 1, standard metric learning models exhibit a critical "semantic blindness." SapiensID can confidently assign high similarity to two face images of different genders or to two whole-body images because they share similar clothing. It relies on entangled visual textures rather than separating identity-preserving attributes from transient noise.

Second, human perception naturally leverages *kinematic continuity*—the consistent, temporal evolution of a subject's structural mechanics (*e.g.*, gait) while in motion. SapiensID, however, operates entirely on static frames. For video tasks such as gait recognition, it averages per-frame features, discarding the discriminative kinematic signatures of human movement.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Face recognition & person reID **(b) Gait recognition**

Figure 1: SapiensID [32], based on face recognition and person reID, mistakenly assigns high similarity to pairs, since it does not consider fixed soft biometrics like gender and heavily relies on dynamic soft biometrics like clothing. Conversely, our approach, SapiensID 2.0, utilizes information extracted by MLLM from images to align fixed soft biometrics with biometric features and disentangle noise to achieve more robust human recognition. Additionally, we extend SapiensID to gait recognition by explicitly modeling the subject's kinematics in videos.

Elevating foundation models to capture semantic awareness and kinematic continuity presents a data scarcity challenge. On the semantic front, large-scale reID datasets like WebBody [32] lack soft biometric annotations such as race, gender, and clothing. They are collected via keyword queries and face similarities, which do not provide semantic labels. On the temporal front, WebBody consists of more than 90% static images. Compiling and labeling a video dataset of comparable scale to learn kinematic signatures from scratch is not feasible.

To overcome the semantic annotation bottleneck, we leverage multimodal large language models (MLLMs, *e.g.*, [5]), which can generate accurate pseudo-ground truths for both fixed and dynamic soft biometrics. As deploying generative MLLMs at inference is computationally infeasible, we transfer their semantic knowledge into a more discriminative feature space. This transfer introduces a structural mismatch: semantic attributes are inherently low-dimensional signals (*e.g.*, age, a scalar, or race, categorical), whereas MLLM embeddings are highly entangled and high-dimensional. Directly aligning these spaces would force the biometric model to absorb irrelevant linguistic artifacts.

To resolve this, we propose a Bipartite Subspace Alignment strategy. For fixed soft biometrics, we introduce Invariant Trait Alignment (ITA), which distills the core variance of stable identity traits into the biometric space. Conversely, to handle dynamic soft biometrics, we introduce Transient Noise Disentanglement (TND), which enforces strict orthogonality against the principal components of transient traits, effectively neutralizing clothing bias.

Simultaneously, to model kinematic continuity without requiring a massive, newly labeled video dataset, we introduce the Kinematic Semantic Attention Head (K-SAH). Rather than training video architectures from scratch, K-SAH harnesses the temporal modeling potential of our image-trained model. By extending the existing semantic attention mechanism across a temporal window, the network explicitly tracks the spatial evolution of the same semantic patches (*e.g.*, body joints) across sequential frames. Crucially, K-SAH is versatile: it exploits kinematic motion patterns when presented with video tracklets, while degenerating to spatial-only processing for static images.

Integrating semantic awareness and kinematic continuity, SapiensID 2.0 serves as a unified human recognition foundation model across face recognition, person reID, and gait recognition. By a unified foundation model, we mean a single model with one set of weights that performs all three tasks without dataset-specific fine-tuning. Prior unified models cover only a subset of these tasks: SapiensID does not address gait recognition, and VILLS [24] does not address face recognition and requires joint video-image training. Extensive experiments demonstrate that SapiensID 2.0 significantly outperforms the previous state of the art on video-based and long-term person reID and on gait recognition, remains competitive on short-term reID, and maintains robust face recognition capabilities.

Our contributions are as follows:

- **Semantic Subspace Alignment via Distillation**: We propose a feature-based knowledge distillation framework for semantic subspace alignment using MLLMs. It employs mathematically distinct pathways—ITA and TND—to simultaneously distill fixed soft biometrics and purge dynamic soft biometric noise from the discriminative feature space.

- **Kinematic Semantic Attention Head (K-SAH)**: We adapt spatial semantic attention to process temporal sequences via tracked structural patches and window attention. It captures kinematic continuity from video without extensive video training, while degenerating to standard static processing for individual images.
- **A Unified Model for Human Recognition**: By integrating semantic awareness and kinematic continuity, SapiensID 2.0 provides a unified foundation model spanning face, reID, and gait recognition, achieving superior performance in generalized recognition tasks.

2 Related Work

Human Recognition. Conventional biometric systems isolate human recognition into specialized modalities, *e.g.*, face recognition [13, 31, 33, 34, 46, 54, 65–67, 73] and person reID [7, 11, 18, 29, 35, 37, 44, 53, 56, 60, 64, 74, 76, 84]. While multi-task frameworks attempt to fuse these domains [38, 39, 55, 80, 81], or explore joint video-image pretraining (*e.g.*, VILLS [24]), they either rely on late-stage decision-level score fusion [81] or require joint video training while omitting face recognition [24]. SapiensID [32] established a unified representation paradigm by training on massive body-centric data, yielding robust, pose-invariant features across face and body. Despite this unification, SapiensID remains semantically agnostic and temporally static—frequently making erroneous geometric matches that contradict obvious human attributes. SapiensID 2.0 extends this foundation by (1) grounding the embedding space in explicit semantics distilled from MLLMs, and (2) extending spatial attention temporally to dynamically track tokens across frames.

Soft Biometrics. Soft biometrics have proven effective [1] in improving face [27, 45, 57] and fingerprint recognition [2, 26] accuracy. However, their reliance on manually annotated attributes [48] severely limits their experimental scale. In clothing-changing person reID [17, 68], disentangling identity from transient appearance relies on clothing labels. Recently, DIFFER [37] addressed this using an adversarial Gradient Reversal Layer, which can be unstable during training. For foundation models, training stability is even more important. In contrast, our framework avoids adversarial training and transfers zero-shot MLLM knowledge purely via a stable, feature-based distillation [59], in the spirit of using language-specified attributes to diagnose and correct attribute reliance in vision models [9]. By employing ITA to directly embed fixed anatomical traits and TND to explicitly orthogonalize dynamic clothing noise, our approach achieves robust semantic disentanglement.

Gait Recognition. Gait recognition identifies individuals by their walking posture, using modalities such as appearance [19, 22, 23, 28, 61, 69], skeletons [16], and LiDAR [50]. Most methods—whether utilizing CNNs [15], transformers [14], or large vision models [71, 72]—require large-scale video datasets for training and inference. Another line of work explores image-based or static-to-dynamic approaches. Early methods compress entire gait cycles into static templates, such as Gait Energy Images (GEIs), though these suffer from severe information loss. Xu *et al*. [63] reconstruct a full gait sequence from a single image during inference; however, training their reconstruction module still requires video data. While general spatio-temporal video transformers (*e.g.*, ViViT [4]) rely on dense grid-based self-attention with quadratic complexity and require video fine-tuning, we demonstrate that a foundation model trained primarily on images can inherently capture temporal dynamics. By adapting the spatial attention mechanism to track semantically anchored keypoint tokens across neighboring frames, our approach captures rich kinematic signatures directly without video retraining.

3 Method

3.1 Preliminaries

SapiensID [32] uses RetinaPatch (RP) to dynamically allocate patches to semantically critical regions (*i.e.*, face, upper torso, whole body). RP generates semantically grounded patches with region-sampled positional embeddings that preserve spatial origin—an inductive bias critical for temporal tracking [82]. These patches pass through the ViT [12] backbone to extract a deep spatial feature map $\mathbf{X} \in \mathbb{R}^{L \times C}$, where L is the number of spatial tokens and C is the feature dimension.

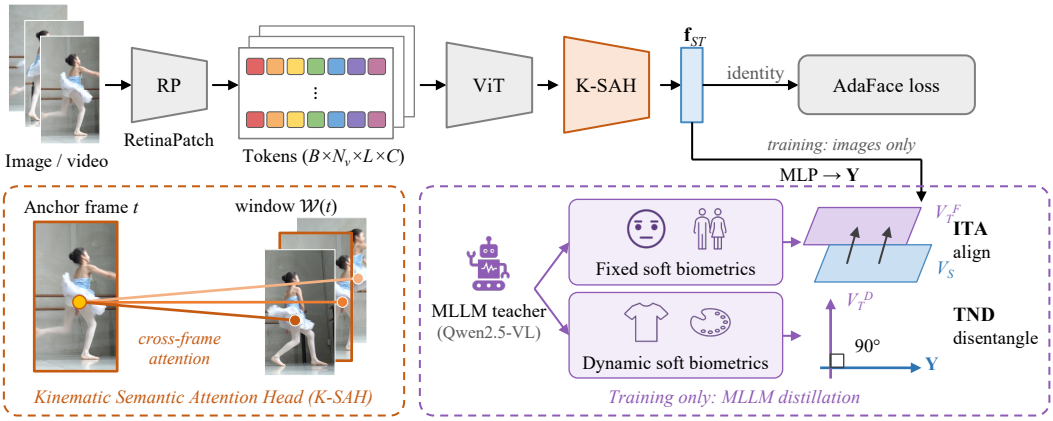


Figure 2: Overall architecture of SapiensID 2.0. RetinaPatch (RP) converts the image into patches based on the landmarks and key points. ViT extracts features from these patches. K-SAH aggregates the features around each joint in images or videos. Additionally, we supervise the training via alignment with fixed soft biometrics and the disentanglement of dynamic soft biometrics.

Following the ViT, the Semantic Attention Head (SAH) aggregates this feature map. SAH generates a semantic query matrix $\mathbf{Q} \in \mathbb{R}^{nK\times C}$ from predicted keypoints using the learned 2D positional embedding grid $\mathbf{P} \in \mathbb{R}^{H\times W\times C}$:

$$\underbrace{\mathbf{Q}}_{nK\times C} = \text{Repeat}_n\big(\text{GridSample}\big(\underbrace{\mathbf{P}}_{H\times W\times C}, \underbrace{\mathbf{C}}_{K\times 2}\big)\big) + \underbrace{\mathbf{O}}_{nK\times C}, \tag{1}$$

where K is the number of keypoints ($K = 19$: body joints from OpenPose [6] and facial landmarks), $\mathbf{C}$ holds the image-specific predicted 2D keypoint coordinates, GridSample$(\cdot)$ bilinearly interpolates $\mathbf{P}$ at each coordinate in $\mathbf{C}$, Repeat$_n$ repeats each sampled embedding n times ($n = 4$ queries per keypoint), and $\mathbf{O}$ is a learned offset matrix that makes the n queries of a keypoint distinct. These queries extract local features via a cross-attention operation over the spatial tokens:

$$\mathbf{H} = \text{Attention}(\mathbf{Q}, \hat{\mathbf{P}}, \mathbf{X}). \tag{2}$$

Here, the key $\hat{\mathbf{P}} \in \mathbb{R}^{L\times C}$ holds the positional embeddings of the L tokens, which RP obtains by resampling $\mathbf{P}$ over each token's region, and $\mathbf{X}$ serves as the value. This allows the network to adaptively pool features around keypoints. As in SapiensID, each row of $\mathbf{H}$ is paired with the token of $\mathbf{X}$ whose positional embedding is closest to the query, queries of undetected keypoints are zeroed, and the result is flattened and projected by two linear layers with batch normalization to the final biometric feature $\mathbf{f} \in \mathbb{R}^{4096}$.

However, SapiensID has two limitations. First, its feature space lacks explicit awareness of fixed soft biometrics and is vulnerable to dynamic soft biometrics. Second, when given a video, it extracts biometric features for each frame and collapses them via averaging. This isolates the spatial attention (Eq. 2) to individual frames and discards the kinematic continuity inherent in temporal dynamics.

3.2 Overall Framework Formulation

We propose SapiensID 2.0 (Fig. 2), a unified architecture driven by explicit soft biometrics and temporal dynamics. Like SapiensID, it uses RP and a ViT backbone to extract structurally grounded features $\mathbf{X}$ from each frame.

Our objective is to extract an enhanced spatio-temporal feature $\mathbf{f}_{ST}$ that satisfies two novel constraints:

1. **Bipartite Semantic Alignment**: The projection of $\mathbf{f}_{ST}$ must be highly correlated with the semantic subspace of *fixed soft biometrics* and strictly orthogonal to the confounding subspace of *dynamic soft biometrics*.
2. **Kinematic Continuity**: The network must explicitly model the trajectory of the semantic queries $\mathbf{Q}$ across time, capturing the continuous kinematic signatures (*e.g.*, gait) that humans use for identification.

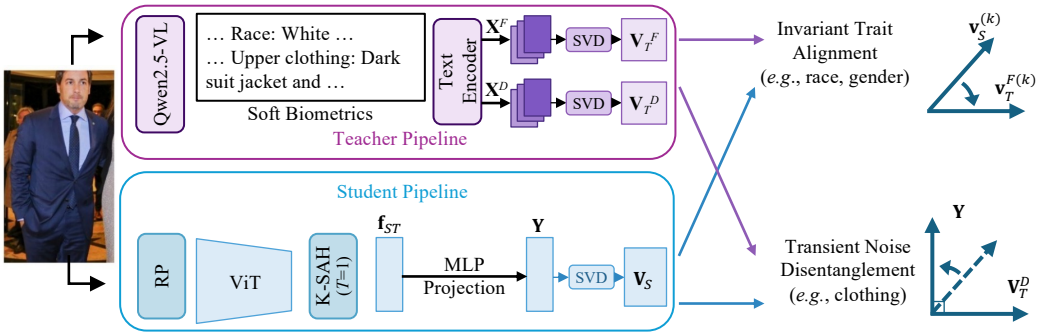


Figure 3: Semantic alignment via distillation. ITA aligns the principal components of the projected student feature $\mathbf{Y}$ with the principal components of fixed soft biometric embeddings in the MLLM space. TND enforces that $\mathbf{Y}$ is orthogonal to those of the confounding factors like clothing, thereby minimizing their influence.

We achieve the first constraint through a novel feature-based distillation module incorporating ITA and TND. To address the second constraint, we replace the static SAH with our K-SAH, which extends cross attention pooling over temporal windows.

3.3 Semantic Alignment via Distillation

As deploying a generative MLLM at the edge is prohibitively expensive, we propose a feature-based cross-modal distillation framework to transfer its semantic understanding into the SapiensID student model (see Fig. 3).

Semantic Subspace Construction. To generate the requisite semantic supervision, we query an MLLM offline on the WebBody4M dataset. It outputs attributes in two groups: Fixed Soft Biometrics (*e.g.*, race, gender) and Dynamic Soft Biometrics (*e.g.*, clothing, hair). These textual descriptions pass through the frozen text encoder, yielding high-dimensional semantic embeddings representing the native hidden spaces: $\mathbf{X}^F$ for fixed soft biometrics and $\mathbf{X}^D$ for dynamic soft biometrics.

To bridge the dimensional gap between the visual metric space and the language space during training, the biometric feature is passed through a shallow Multi-Layer Perceptron (MLP), yielding the projected student feature $\mathbf{Y}$ (training uses static images, so this is the single-frame case of $\mathbf{f}_{ST}$). We force this feature to align with invariant traits while orthogonalizing against transient noise. This MLP and the teacher text encoder are used only for training losses and discarded at inference, so SapiensID 2.0 runs the same network as SapiensID in a single forward pass with no added inference cost.

Invariant Trait Alignment (ITA). Fixed soft biometrics are low-dimensional signals representing stable identity markers. ITA aligns the orthogonal basis vectors of highest semantic variance in the teacher's fixed subspace with the SapiensID feature space. Computing the empirical covariance of $\mathbf{X}^F$ and applying Singular Value Decomposition (SVD), we extract the top D_F eigenvectors, represented by the matrix $\mathbf{V}_T^F \in \mathbb{R}^{C \times D_F}$. A parallel decomposition on the projected student feature $\mathbf{Y}$ yields its principal components matrix, $\mathbf{V}_S$. The alignment loss is:

$$\mathcal{L}_{ITA} = \sum_{k=1}^{D_F} \left(1 - \left\langle \mathbf{v}_T^{F(k)}, \mathbf{v}_S^{(k)} \right\rangle^2\right) + \lambda_P \left\| \mathbf{X}^F \mathbf{V}_T^F - \mathbf{Y}\mathbf{V}_S \right\|_F^2, \tag{3}$$

where $\mathbf{v}_T^{F(k)}$ and $\mathbf{v}_S^{(k)}$ denote the k-th eigenvectors (columns) of $\mathbf{V}_T^F$ and $\mathbf{V}_S$, respectively, with $D_F = 6$ corresponding to the fixed soft-biometric categories (gender, race, age, skin tone, body build, and facial structure). The first term aligns the geometric axes of the latent spaces by maximizing the squared cosine similarity between teacher and student principal components; the second aligns the data distributions along these axes by minimizing the Mean Squared Error between the projected semantic and visual coordinates. Together, they transfer invariant biometric traits while filtering out the MLLM's contextual noise.

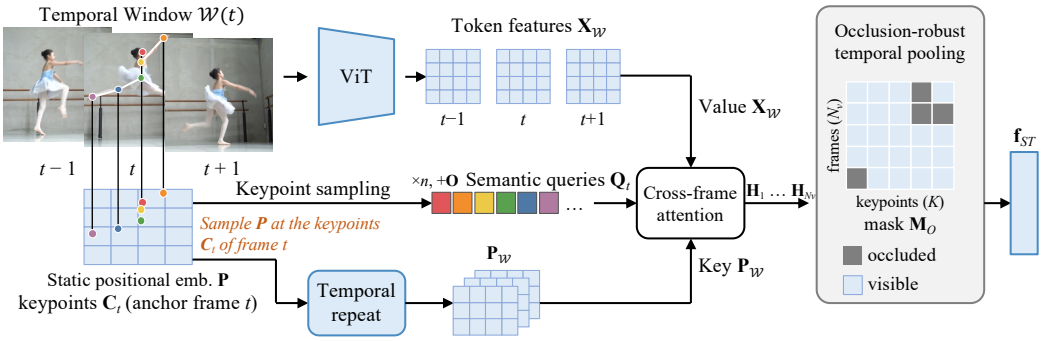


Figure 4: Illustration of K-SAH. It extends static semantic attention to model local temporal dynamics over a temporal window $\mathcal{W}$. Semantic queries $\mathbf{Q}_t$ for the anchor frame t are derived via grid sampling from static embeddings, preserving spatial priors. Cross-frame attention between the queries $\mathbf{Q}_t$ and dense spatiotemporal key $\mathbf{P}_\mathcal{W}$ and value $\mathbf{X}_\mathcal{W}$ volumes allows precise tracking of geometric trajectories. Occlusion-Robust Temporal Pooling dynamically aggregates intermediate features $\mathbf{H}_t$ by masking occluded keypoints using an explicit mask $\mathbf{M}_O$, producing the final flattened spatiotemporal feature $\mathbf{f}_{ST}$.

Transient Noise Disentanglement (TND). Dynamic soft biometrics must be explicitly decoupled from the identity representation to prevent reliance on transient attributes such as identical outfits. Applying SVD to the empirical covariance of $\mathbf{X}^D$, we extract its top D_D principal components, forming the matrix $\mathbf{V}_T^D \in \mathbb{R}^{C \times D_D}$, which defines the geometric subspace of transient noise ($D_D = 5$ corresponding to hair, facial hair, accessories, upper clothing, and lower clothing). We force $\mathbf{Y}$ to be orthogonal to this subspace via a projection penalty:

$$\mathcal{L}_{TND} = \left\| \mathbf{Y}\mathbf{V}_T^D \right\|_F^2 . \tag{4}$$

Relational Topology Preservation. To complement ITA, we preserve the relational topology of instances within a training batch by comparing the Gram matrices of the teacher and student features:

$$\mathcal{L}_{FR} = \frac{1}{N^2} \| \mathbf{X}^F (\mathbf{X}^F)^\top - \mathbf{Y}\mathbf{Y}^\top \|_F^2 . \tag{5}$$

where N is the number of images in the batch. This ensures that individuals with similar fixed soft biometrics in the language space are proportionally close in the metric space.

3.4 Kinematic Semantic Attention Head (K-SAH)

Extending human recognition to the temporal domain is hindered by data availability: static image datasets (*e.g.*, WebBody4M [32]) vastly exceed video datasets in scale, diversity, and annotation quality, so training parameter-heavy spatiotemporal models on limited video data leads to severe overfitting and the loss of pre-trained spatial priors.

Instead of building a separate video architecture, K-SAH lifts the pre-trained image model into the temporal domain. Because SapiensID's spatial tokens encode precise keypoint locations via the RP positional embeddings, we can natively track these trajectories over time, capturing *kinematic continuity* simply by expanding the attention window.

This design guarantees checkpoint compatibility: the pre-trained static SAH weights transfer directly to K-SAH, and SAH is the special case of K-SAH for a single image ($N_v = T = 1$).

Temporal Windowing and Spatio-Temporal Embeddings. Given a video tracklet of N_v frames yielding spatial feature tensor $\mathbf{X} \in \mathbb{R}^{B \times N_v \times L \times C}$ (where B is the batch size), we explicitly model the local temporal dynamics without flattening the entire video sequence, which would cause quadratic attention bottlenecks. For each frame t, we define a localized temporal window of T frames centered on t, $\mathcal{W}(t) = \{t - \lfloor T/2 \rfloor, \dots, t + \lfloor T/2 \rfloor\}$, truncated at the tracklet boundaries.

We add no temporal positional embedding: the keys of a window are the 2D token positional embeddings of its frames, $\mathbf{P}_\mathcal{W} = [\hat{\mathbf{P}}_{t'}]_{t' \in \mathcal{W}(t)} \in \mathbb{R}^{|\mathcal{W}| L \times C}$, where $\hat{\mathbf{P}}_{t'}$ are the token positional

embeddings of frame t'. Because RP samples these embeddings relative to each frame's face and body regions, the same body location receives a similar embedding in every frame, and the spatial knowledge from WebBody4M image pre-training is preserved.

Multi-Frame Semantic Attention. Within K-SAH, the semantic query $\mathbf{Q}_t$ is generated for the anchor frame t via grid sampling of $\mathbf{P}$ using its frame-level keypoint predictions $\mathbf{C}_t$, similar to Eq. 1:

$$\mathbf{Q}_t = \text{Repeat}_n\big(\text{GridSample}\left(\mathbf{P}, \mathbf{C}_t\right)\big) + \mathbf{O}. \tag{6}$$

However, rather than attending solely to the spatial tokens of frame t (as in Eq. 2), the query attends to the concatenated spatial features $\mathbf{X}_{\mathcal{W}} \in \mathbb{R}^{|\mathcal{W}|L \times C}$ and their corresponding 2D positional embeddings $\mathbf{P}_{\mathcal{W}}$ within the temporal window. The multi-frame cross-attention yields an intermediate temporal feature matrix for all keypoints at frame t:

$$\mathbf{H}_t = \text{Attention}\left(\mathbf{Q}_t, \mathbf{P}_{\mathcal{W}}, \mathbf{X}_{\mathcal{W}}\right). \tag{7}$$

Unlike global spatio-temporal attention over the tracklet, which scales quadratically as $\mathcal{O}(N_v^2 L^2 C)$, K-SAH restricts cross-attention strictly along the K keypoint trajectories, reducing complexity to linear scaling $\mathcal{O}(T \cdot K \cdot L \cdot C)$ with respect to window size T. Being compatible with the pre-trained static weights, it replaces isolated spatial pooling without video retraining and lets attention track specific semantic patches (*e.g.*, a swinging arm or stepping foot) across neighboring frames, capturing continuous kinematics.

Occlusion-Robust Temporal Pooling. In unconstrained video, keypoints frequently disappear and reappear due to occlusion; K-SAH handles this via masked temporal pooling before flattening.

Let $\mathbf{M}_O \in \{0,1\}^{B \times N_v \times K}$ be a boolean mask indicating whether keypoint $k \in \{1, \ldots, K\}$ is missing or severely occluded in frame t. $\mathbf{M}_O[t,k] = 1$ when keypoint k is not detected in frame t, *i.e.*, when its OpenPose [6] confidence falls below the peak-detection threshold of 0.1. Let $\mathbf{h}_{t,k}$ denote the features of keypoint k (its n query rows) in $\mathbf{H}_t$. We pool the intermediate multi-frame features $\mathbf{h}_{t,k}$ across the temporal dimension, explicitly disregarding occluded patches:

$$\bar{\mathbf{h}}_k = \frac{\sum_{t=1}^{N_v} \mathbf{h}_{t,k} \cdot (1 - \mathbf{M}_O[t,k])}{\max\left(1, \sum_{t=1}^{N_v} (1 - \mathbf{M}_O[t,k])\right)}. \tag{8}$$

This prevents occluded frames from corrupting the aggregate; the pooled features are flattened into the final spatiotemporal feature $\mathbf{f}_{ST}$.

4 Experiments

Experiment Setup. Unless otherwise noted, our experimental setup, including the ViT backbone, RetinaPatch extraction, and training hyperparameters, strictly follows SapiensID [32]. For the semantic distillation modules, we utilize `Qwen2.5-VL-7B-Instruct` [5] to extract the semantic subspaces offline, retaining the top principal components for both ITA ($D_F = 6$) and TND ($D_D = 5$). Training uses static images only; for video-based tasks, K-SAH is applied at inference with a temporal window of $T = 3$ frames. We train a single model on WebBody, constructed exactly as in SapiensID [32] (including the additional body reID training sets that complement the face-similarity-labeled web images), and evaluate it directly in a zero-shot manner on all testing benchmarks without dataset-specific fine-tuning. For video-based evaluation on CCPG, we evaluate on RGB video sequences, extracting kinematic trajectories via K-SAH. Because baseline models are bound to their input modalities (gait models strictly require video data like CCGR, while appearance models are trained on static corpora), all models are evaluated zero-shot on the target test sets for a fair comparison of out-of-distribution generalization (see Tab. 6 in the supplementary material for baseline pretraining metadata). We use the same checkpoint for each baseline model.

Video-Based Person ReID and Gait Recognition. Table 1 evaluates SapiensID 2.0 on video-based person reID (CCVID [17]) and gait recognition (CCPG [36]). The results clearly illustrate the divergence between models that overfit to domain-specific cues and those that learn invariant representations. On CCVID, while video-specialized models like BiggerGait [71] achieve strong overall accuracy, SapiensID 2.0 achieves top performance (95.66% General Rank-1). More importantly, it

Table 1: Generalization comparison with SoTA ReID and gait recognition models. The evaluation is split into video-based person reID (CCVID) and gait recognition (CCPG). For CCPG, we report the rank-1 accuracy under four evaluation protocols. SapiensID 2.0 leverages semantic and kinematic priors to achieve SoTA performance.

Method	CCVID (General)		CCVID (CC)	
	top1	mAP	top1	mAP
CAL [17]	74.83	29.43	73.89	26.65
CLIP3DReID [41]	77.28	30.01	76.28	26.69
SOLIDER [8]	40.27	36.56	39.61	35.48
HAP [75]	54.74	45.14	52.37	41.33
BiggerGait [71]	85.13	83.38	84.53	**82.59**
SapiensID [32]	92.57	77.82	**88.72**	72.22
Ours	**95.66**	**85.00**	88.63	78.05

Method	CCPG [36]			
	CL	UP	DN	BG
CLIP3DReID [41]	15.7	25.6	24.5	60.3
GaitBase [15]	20.8	31.3	38.3	70.2
DeepGaitV2 [14]	23.0	37.1	36.5	69.9
BigGait [72]	22.9	42.2	25.0	80.5
BiggerGait [71]	23.2	44.5	29.5	86.7
SapiensID [32]	33.9	45.0	75.2	90.2
Ours	**60.8**	**60.8**	**86.4**	**91.6**

Table 2: Generalization comparison with SoTA ReID models across short-term and long-term settings. "Long-term" refers to clothing change (CC) protocols of LTCC and PRCC datasets, while "short-term" uses the same clothing (SC) protocol. For other datasets, the data capture characteristics define short or long-term conditions. While texture-biased models excel marginally on short-term benchmarks, SapiensID 2.0 achieves much superior robustness and generalization in long-term scenarios.

Method	Avg	LTCC (General) [51]		PRCC (SC) [68]		Market1501 [77]		MSMT17 [62]	
		top1	mAP	top1	mAP	top1	mAP	top1	mAP
CAL [17]	49.09	72.41	38.12	99.54	99.01	43.65	21.03	14.48	4.44
CLIP3DReID [41]	50.20	75.66	**45.15**	99.43	96.43	41.66	20.33	17.45	5.50
SOLIDER [8]	**71.45**	73.83	36.28	99.51	**99.53**	**97.03**	**94.04**	48.64	22.77
HAP [75]	70.78	73.02	35.97	99.30	98.45	96.23	92.20	48.01	23.02
SapiensID [32]	70.01	72.01	34.56	**100.00**	98.79	88.18	68.26	67.25	**31.02**
SapiensID 2.0 (Ours)	71.33	**77.08**	42.39	**100.00**	99.01	86.94	67.15	**67.72**	30.32

(a) Short-Term ReID

Method	Avg	LTCC (CC) [51]		PRCC (CC) [68]		CCDA [40]		Celeb-ReID [25]	
		top1	mAP	top1	mAP	top1	mAP	top1	mAP
CAL [17]	24.26	33.16	16.27	45.39	45.42	3.74	9.14	37.11	3.81
CLIP3DReID [41]	24.93	41.84	22.58	40.81	38.38	4.31	10.18	37.31	4.02
SOLIDER [8]	21.66	25.00	12.18	26.87	32.12	8.62	16.48	46.37	5.66
HAP [75]	22.66	24.74	11.71	33.90	37.00	8.30	16.02	44.38	5.20
SapiensID [32]	62.77	42.35	17.79	78.75	72.60	61.84	69.08	92.80	66.92
SapiensID 2.0 (Ours)	**68.82**	**50.26**	**24.28**	**87.19**	**81.32**	**66.97**	**73.06**	**93.64**	**73.87**

(b) Long-Term ReID

demonstrates robustness in the challenging clothing-change (CC) protocol, boosting SapiensID's mAP by $+5.83\%$ (from 72.22% to 78.05%). This suggests that explicitly penalizing dynamic visual traits (via TND) encourages the model to synthesize a reliable temporal identity rather than simply tracking an outfit across frames.

The performance on CCPG further exposes the generalization bottlenecks of specialized architectures. Models trained on ReID data (*e.g.*, CLIP3DReID [41]) inherently struggle to capture identity information from gait data. Conversely, the poor cross-dataset generalization of gait models indicates they likely overfit to their training domains (*e.g.*, BiggerGait reaches only 23.2% Rank-1 on CCPG CL). By leveraging K-SAH to explicitly track the temporal displacement of spatial patches through keypoints, SapiensID 2.0 extracts robust kinematic signatures, substantially improving performance on CCPG CL (33.9% to 60.8%) without requiring massive video retraining.

Image-Based Person ReID. We evaluate SapiensID 2.0 on image-based person reID benchmarks where the improvements come from the semantic alignment with MLLM. The distinction between short-term (same clothing) and long-term (clothing change) protocols reveals a critical flaw in previous methods: short-term datasets inherently reward models that overfit to transient appearance

Table 3: 1:1 Face verification accuracy across face recognition benchmarks. These results confirm that SapiensID 2.0 fully preserves SoTA facial recognition capabilities.

Method	AdaFace [30]	SapiensID [32]	SapiensID 2.0 (Ours)
LFW [21]	99.82	99.82	**99.85**
CPLFW [78]	95.12	94.85	**95.28**
CFP-FP [49]	**99.19**	98.74	99.14
CALFW [79]	**96.07**	95.78	95.93
AgeDB [43]	97.97	97.33	**98.07**
Avg	97.63	97.30	**97.65**

Table 4: Component-wise ablation of SapiensID 2.0 on the CCVID dataset. "Gen" denotes the general evaluation protocol, "CC" denotes the clothing-change protocol. The last three rows vary the K-SAH temporal window T (default T=3). The results confirm that combining these components yields highly synergistic performance gains.

ITA	TND	K-SAH	CCVID (Gen)		CCVID (CC)	
			top1	mAP	top1	mAP
-	-	-	92.57	77.82	88.72	72.22
✓	-	-	92.70	80.70	87.97	73.58
✓	✓	-	93.82	80.70	**89.97**	75.58
✓	✓	T=3	95.66	**85.00**	88.63	**78.05**
✓	✓	T=5	95.65	84.95	88.52	77.95
✓	✓	T=7	**95.70**	84.90	88.50	77.87

cues. As shown in Tab. 2a, models like SOLIDER perform exceptionally well on short-term tasks (averaging 71.45% accuracy), which we attribute to exploiting texture and clothing.

This is supported by SOLIDER's average collapsing to 21.66% in long-term scenarios where subjects change outfits (Tab. 2b). In contrast, SapiensID 2.0 deliberately penalizes reliance on these transient cues via the TND objective. While this occasionally causes our model to marginally underperform SOLIDER on short-term benchmarks (averaging 71.33%), it fundamentally anchors the latent space to invariant anatomical traits (via ITA). Consequently, SapiensID 2.0 is consistently robust across long-term protocols, elevating the baseline average from 62.77% to 68.82%.

Face Recognition. To verify that optimizing for body-centric soft biometrics and temporal kinematics does not induce catastrophic forgetting of fine-grained facial features, we evaluate SapiensID 2.0 across five standard face recognition benchmarks. As shown in Tab. 3, SapiensID 2.0 fully preserves, and slightly exceeds, SoTA face verification accuracy (*e.g.*, 99.85% on LFW vs. AdaFace-ViT's 99.82%). We hypothesize this is due to ITA: the invariant soft biometrics extracted from the MLLM (*e.g.*, gender) correlate with structural facial features, so the semantic alignment avoids forgetting fine-grained facial priors.

Ablation Studies. We ablate our modules on CCVID [17], which provides video tracklets and separate protocols for general (short-term) and clothing-change (long-term) scenarios. Tab. 4 reports the progressive integration of our modules over the semantically agnostic, static baseline. Introducing ITA provides an immediate boost to both general and clothing-change mAP (*e.g.*, General mAP jumps from 77.82% to 80.70%), suggesting that anchoring the geometric space to stable semantic priors (*e.g.*, body proportions, gender) enhances overall discriminability. The addition of TND yields a particularly pronounced improvement on the CCVID (CC) split, increasing CC mAP by a further $+2.00\%$ and CC top-1 by $+2.00\%$. By enforcing orthogonality against dynamic noise, TND prevents overfitting to identical outfits, mitigating the baseline's "semantic blindness." Swapping SAH for K-SAH to model temporal dynamics (fourth row) drives the final push to SoTA metrics, improving General mAP from 80.70% to 85.00% and CC mAP from 75.58% to 78.05%, at the cost of 1.34% CC top-1. The last rows of Tab. 4 vary the K-SAH temporal window: a short window already captures the gain, and performance saturates at $T = 3$ (78.05% CC mAP), so larger windows only add computation (see the supplementary material for the comparison with frame averaging).

Distillation Design. Offline probes on frozen SapiensID features (with CLIP-encoded text targets; supplementary material) support two design choices. First, distilling from re-encoded MLLM text rather than raw MLLM visual hidden states yields the same alignment error with $4\times$ lower residual clothing leakage (R^2 $0.080 \rightarrow 0.020$), because the text retains only the queried attributes. Second, directly regressing the fixed-trait embeddings reaches the same alignment error as ITA+TND but leaves a component along the dynamic subspace (orthogonality error 0.026 *vs.* 2.8×10^{-6}).

Efficiency. As MLLM features are extracted offline, a training iteration increases only from 1.77 s (baseline) to 2.06 s. During inference, the only added computational cost is the temporal attention in the K-SAH module. With K-SAH (T=3), SapiensID 2.0 requires 68.7 GFLOPs per frame and runs at

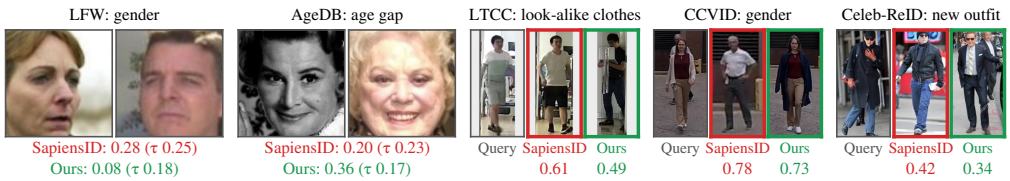


Figure 5: Cases where SapiensID fails and SapiensID 2.0 (Ours) succeeds. Left: face verification of a different-identity pair (LFW) and a same-identity pair (AgeDB); each model's cosine similarity is followed by its verification threshold τ on that dataset (a pair is accepted as the same identity if the similarity is at least τ); red: wrong decision, green: correct. Right: query and the rank-1 match of SapiensID (red, wrong identity) and of Ours (green, correct), with cosine similarity to the query. SapiensID confuses gender and look-alike clothes and misses the same person across an age gap or an outfit change.

23.5 ms (42.6 FPS) on an NVIDIA H100 (fp16, batch size 1, excluding keypoint detection), a 224$\times$ higher throughput than querying an edge MLLM such as MobileVLM V2 [10] for the attributes (see the supplementary material).

Visualizations & Discussion. To qualitatively validate our semantic distillation, Fig. 5 illustrates cases where the baseline SapiensID fails but SapiensID 2.0 succeeds, across faces and bodies and across indoor, surveillance and in-the-wild images. The baseline model suffers from "semantic blindness": it assigns high similarity to different individuals who differ in gender (fixed soft biometrics) or share look-alike clothing (dynamic soft biometrics), and it misses the same person across an age gap or an outfit change. SapiensID 2.0 corrects both kinds of error, in line with human perception.

5 Limitations

The semantic awareness of SapiensID 2.0 is bounded by the teacher MLLM, whose soft-biometric labels are not error-free: it can hallucinate attributes under occlusion or unusual poses, or attribute a nearby person's hair or clothing to the subject. WebBody4M is collected via single-subject name queries, which limits this contamination, and because ITA and TND act on corpus-level principal components, isolated labeling errors have a limited effect; systematic MLLM biases (*e.g.*, in perceived race or age), however, can propagate to the student. Suppressing clothing improves long-term generalization but costs a little accuracy on short-term benchmarks (*e.g.*, Market1501), where clothing is a valid retrieval cue (see the supplementary material for failure cases). Extracting the semantic subspaces requires a one-time pass of the MLLM over the training set; the MLLM and the distillation projections are discarded after training, so inference cost is unaffected.

6 Conclusion

We present SapiensID 2.0, a unified foundation model that grounds human recognition in explicit semantic awareness and kinematic continuity. We introduce a semantic distillation framework utilizing Bipartite Subspace Alignment. This strategy leverages ITA and TND to anchor identity to stable anatomical traits while explicitly decoupling transient artifacts. Furthermore, our K-SAH extends spatial attention to track structural patches over time, capturing continuous motion signatures without the prohibitive overhead of video-level retraining. Extensive experiments demonstrate that SapiensID 2.0 achieves state-of-the-art performance across image- and video-based person re-identification and gait recognition, while preserving robust face recognition capabilities. Beyond human recognition, our results suggest two broader findings: textual MLLM descriptions can decompose a visual representation into invariant and transient subspaces, offering a general recipe for robustness to nuisance factors, and a strong image model with spatially grounded tokens can be lifted to sequence tasks via keypoint-anchored attention without video training.

Acknowledgements and Disclosures. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is autho-

rized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The authors declare no competing interests.

References

[1] Abdelgader Abdelwhab and Serestina Viriri. "A survey on soft biometrics for human identification". In: *Machine Learning and Biometrics*. IntechOpen, 2018.

[2] Heikki Ailisto, Elena Vildjiounaite, Mikko Lindholm, Satu-Marja Mäkelä, and Johannes Peltola. "Soft biometrics—combining body weight and fat measurements with fingerprint biometrics". In: *PRL* (2006).

[3] Daniel Arkushin, Bar Cohen, Shmuel Peleg, and Ohad Fried. "GEFF: Improving Any Clothes-Changing Person ReID Model Using Gallery Enrichment With Face Features". In: *WACV*. 2024.

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. "ViViT: A Video Vision Transformer". In: *ICCV*. 2021.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. "Qwen2.5-VL Technical Report". In: *arXiv preprint arXiv:2502.13923* (2025).

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *TPAMI* (2021).

[7] Keqi Chen, Vinkle Srivastav, Didier Mutter, and Nicolas Padoy. "Learning from Synchronization: Self-Supervised Uncalibrated Multi-View Person Association in Challenging Scenes". In: *CVPR*. 2025.

[8] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. "Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks". In: *CVPR*. 2023.

[9] Xuanbai Chen, Xiang Xu, Zhihua Li, Tianchen Zhao, Pietro Perona, Qin Zhang, and Yifan Xing. "Model Diagnosis and Correction via Linguistic and Implicit Attribute Editing". In: *CVPR*. 2025.

[10] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. "MobileVLM V2: Faster and Stronger Baseline for Vision Language Model". In: *arXiv preprint arXiv:2402.03766* (2024).

[11] Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. "DKC: Differentiated Knowledge Consolidation for Cloth-Hybrid Lifelong Person Re-identification". In: *CVPR*. 2025.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*. 2021.

[13] Saeed Ebrahimi, Sahar Rahimi, Ali Dabouei, Srinjoy Das, Jeremy M Dawson, and Nasser M Nasrabadi. "GIF: Generative Inspiration for Face Recognition at Scale". In: *CVPR*. 2025.

[14] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. "Exploring Deep Models for Practical Gait Recognition". In: *arXiv preprint arXiv:2303.03301* (2023).

[15] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. "OpenGait: Revisiting Gait Recognition Towards Better Practicality". In: *CVPR*. 2023.

[16] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. "Skeletongait: Gait recognition using skeleton maps". In: *AAAI*. 2024.

[17] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. "Clothes-changing person re-identification with rgb modality only". In: *CVPR*. 2022.

[18] Ruiyang Ha, Songyi Jiang, Bin Li, Bikang Pan, Yihang Zhu, Junjie Zhang, Xiatian Zhu, Shaogang Gong, and Jingya Wang. "Multi-modal Multi-platform Person Re-Identification: Benchmark and Method". In: *ICCV*. 2025.

[19] Gavriel Habib, Noa Barzilay, Or Shimshi, Rami Ben-Ari, and Nir Darshan. "Cargait: Cross-attention based re-ranking for gait recognition". In: *ICCV*. 2025.

[20] Bilal Hassan, Ebroul Izquierdo, and Tomas Piatrik. "Soft biometrics: A survey: Benchmark analysis, open challenges and recommendations". In: *MTAP* (2024).

[21] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments". In: *ECCVW*. 2008.

[22] Panjian Huang, Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. "Vocabulary-Guided Gait Recognition". In: *NeurIPS*. 2025.

[23] Panjian Huang, Saihui Hou, Junzhou Huang, and Yongzhen Huang. "Learning A Unified Template for Gait Recognition". In: *ICCV*. 2025.

[24] Siyuan Huang, Ram Prabhakar Kathirvel, Yuxiang Guo, Rama Chellappa, and Cheng Peng. "VILLS : Video-Image Learning to Learn Semantics for Person Re-Identification". In: *WACV*. 2025.

[25] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. "Celebrities-ReID: A benchmark for clothes variation in long-term person re-identification". In: *IJCNN*. 2019.

[26] Anil K Jain, Karthik Nandakumar, Xiaoguang Lu, and Unsang Park. "Integrating faces, fingerprints, and soft biometric traits for user recognition". In: *BioAW*. 2004.

[27] Anil K Jain and Unsang Park. "Facial marks: Soft biometric for face recognition". In: *ICIP*. 2009.

[28] Dongyang Jin, Chao Fan, Jingzhe Ma, Jingkai Zhou, Weihua Chen, and Shiqi Yu. "On Denoising Walking Videos for Gait Recognition". In: *CVPR*. 2025.

[29] Wajahat Khalid, Bin Liu, Xulin Li, Muhammad Waqas, and Muhammad Sher Afgan. "Bridging the Sky and Ground: Towards View-Invariant Feature Learning for Aerial-Ground Person Re-Identification". In: *ICCV*. 2025.

[30] Minchul Kim, Anil K Jain, and Xiaoming Liu. "Adaface: Quality adaptive margin for face recognition". In: *CVPR*. 2022.

[31] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. "KeyPoint Relative Position Encoding for Face Recognition". In: *CVPR*. 2024.

[32] Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. "Sapiensid: Foundation for human recognition". In: *CVPR*. 2025.

[33] Minsoo Kim, Min-Cheol Sagong, Gi Pyo Nam, Junghyun Cho, and Ig-Jae Kim. "VIGFace: Virtual Identity Generation for Privacy-Free Face Recognition Dataset". In: *ICCV*. 2025.

[34] Sunpill Kim, Seunghun Paik, Chanwoo Hwang, Dongsoo Kim, Junbum Shin, and Jae Hong Seo. "IDFace: Face Template Protection for Efficient and Secure Identification". In: *CVPR*. 2025.

[35] Longhua Li, Lei Qi, and Xin Geng. "One-Shot Knowledge Transfer for Scalable Person Re-Identification". In: *ICCV*. 2025.

[36] Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. "An In-Depth Exploration of Person Re-Identification and Gait Recognition in Cloth-Changing Conditions". In: *CVPR*. 2023.

[37] Xin Liang and Yogesh S Rawat. "DIFFER: Disentangling Identity Features via Semantic Cues for Clothes-Changing Person Re-ID". In: *CVPR*. 2025.

[38] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, et al. "Farsight: A physics-driven whole-body biometric system at large distance and altitude". In: *WACV*. 2024.

[39] Feng Liu, Nicholas Chimitt, Lanqing Guo, Jitesh Jain, Aditya Kane, Minchul Kim, Wes Robbins, Yiyang Su, Dingqiang Ye, Xingguang Zhang, et al. "Person Recognition at Altitude and Range: Fusion of Face, Body Shape and Gait". In: *arXiv preprint arXiv:2505.04616* (2025).

[40] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. "Learning Clothing and Pose Invariant 3D Shape Representation for Long-Term Person Re-Identification". In: *ICCV*. 2023.

[41] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. "Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation". In: *CVPR*. 2024.

[42] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. "Biometrics recognition using deep learning: A survey". In: *Artificial Intelligence Review* (2023).

[43] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. "Agedb: the first manually collected, in-the-wild age database". In: *CVPRW*. 2017.

[44] Zhiqi Pang, Junjie Wang, Lingling Zhao, and Chunyu Wang. "Identity-Clothing Similarity Modeling for Unsupervised Clothing Change Person Re-Identification". In: *CVPR*. 2025.

[45] Unsang Park and Anil K Jain. "Face matching and retrieval using soft biometrics". In: *TIFS* (2010).

[46] Zhengyuan Peng, Jianqing Xu, Yuge Huang, Jinkun Hao, Shouhong Ding, Zhizhong Zhang, Xin Tan, and Lizhuang Ma. "Stylized-Face: A Million-level Stylized Face Dataset for Face Recognition". In: *ICCV*. 2025.

[47] 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: *ICML*. 2021.

[48] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. "Deep Imbalanced Attribute Classification Using Visual Attention Aggregation". In: *ECCV*. 2018.

[49] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. "Frontal to profile face verification in the wild". In: *WACV*. 2016.

[50] Chuanfu Shen, Rui Wang, Lixin Duan, and Shiqi Yu. "LidarGait++: Learning Local Features and Size Awareness from LiDAR Point Clouds for 3D Gait Recognition". In: *CVPR*. 2025.

[51] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. "Large-Scale Spatio-Temporal Person Re-identification: Algorithms and Benchmark". In: *TCSVT* (2021).

[52] Nyle Siddiqui, Florinel Alin Croitoru, Gaurav Kumar Nayak, Radu Tudor Ionescu, and Mubarak Shah. "DLCR: A Generative Data Expansion Framework via Diffusion for Clothes-Changing Person Re-ID". In: *WACV*. 2025.

[53] Woojung Son, Yoonki Cho, Guoyuan An, Chanmi Lee, and Sung-Eui Yoon. "Towards Robustness of Person Search against Corruptions". In: *ICCV*. 2025.

[54] Yiyang Su, Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. "Open-Set Biometrics: Beyond Good Closed-Set Models". In: *ECCV*. 2024.

[55] Yiyang Su, Minchul Kim, Jie Zhu, Christopher Perry, Feng Liu, Anil Jain, and Xiaoming Liu. "LocalScore: Local Density-Aware Similarity Scoring for Biometrics". In: *arXiv preprint arXiv:2602.01012* (2026).

[56] Yiyang Su, Yunping Shi, Feng Liu, and Xiaoming Liu. "Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid". In: *ICCV*. 2025.

[57] Pedro Tome, Ruben Vera-Rodriguez, Julian Fierrez, and Javier Ortega-Garcia. "Facial soft biometric features for forensic face recognition". In: *Forensic Science International* (2015).

[58] Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. "FastVLM: Efficient Vision Encoding for Vision Language Models". In: *CVPR*. 2025.

[59] Hong-Ji Wang, Xiang Xu, Bao-Min Xu, Shuang-Yuan Yu, and Quan-Xin Wang. "Student-Centric Network Learning for Improved Knowledge Transfer". In: *JISE* (2020).

[60] Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. "SeCap: Self-Calibrating and Adaptive Prompts for Cross-view Person Re-Identification in Aerial-Ground Networks". In: *CVPR*. 2025.

[61] Zengbin Wang, Saihui Hou, Junjie Li, Xu Liu, Chunshui Cao, Yongzhen Huang, Siye Wang, and Man Zhang. "Gait-X: Exploring X modality for Generalized Gait Recognition". In: *ICCV*. 2025.

[62] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. "Person transfer GAN to bridge domain gap for person re-identification". In: *CVPR*. 2018.

[63] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. "Gait recognition from a single image using a phase-aware gait cycle reconstruction network". In: *ECCV*. 2020.

[64] Kunlun Xu, Fan Zhuo, Jiangmeng Li, Xu Zou, and Jiahuan Zhou. "Self-Reinforcing Prototype Evolution with Dual-Knowledge Cooperation for Semi-Supervised Lifelong Person Re-Identification". In: *ICCV*. 2025.

[65] Xiang Xu. "Face Recognition in the Presence of Variance in Pose, Expression, and Occlusions". PhD thesis. University of Houston, 2019.

[66] Xiang Xu, Pengfei Dou, Ha A. Le, and Ioannis A. Kakadiaris. “When 3D-Aided 2D Face Recognition Meets Deep Learning: An Extended UR2D for Pose-Invariant Face Recognition”. In: *arXiv preprint arXiv:1709.06532* (2017).

[67] Xiang Xu, Ha Le, Pengfei Dou, Yuhang Wu, and Ioannis A. Kakadiaris. “Evaluation of a 3D-Aided Pose-Invariant 2D Face Recognition System”. In: *IJCB*. 2017.

[68] Qize Yang, Ancong Wu, and Wei-Shi Zheng. “Person re-identification by contour sketch under moderate clothing change”. In: *TPAMI* (2019).

[69] Shaopeng Yang, Jilong Wang, Saihui Hou, Xu Liu, Chunshui Cao, Liang Wang, and Yongzhen Huang. “Bridging gait recognition and large language models sequence modeling”. In: *CVPR*. 2025.

[70] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. “Good is Bad: Causality Inspired Cloth-debiasing for Cloth-changing Person Re-identification”. In: *CVPR*. 2023.

[71] Dingqiang Ye, Chao Fan, Zhanbo Huang, Chengwen Luo, Jianqiang Li, Shiqi Yu, and Xiaoming Liu. “Biggergait: Unlocking gait recognition with layer-wise representations from large vision models”. In: *NeurIPS*. 2025.

[72] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. “BigGait: Learning Gait Representation You Want by Large Vision Models”. In: *CVPR*. 2024.

[73] Jinghan You, Shanglin Li, Yuanrui Sun, Jiangchuan Wei, Mingyu Guo, Chao Feng, and Jiao Ran. “LVFace: Progressive Cluster Optimization for Large Vision Models in Face Recognition”. In: *ICCV*. 2025.

[74] Chao Yuan, Guiwei Zhang, Changxiao Ma, Tianyi Zhang, and Guanglin Niu. “From poses to identity: Training-free person re-identification via feature centralization”. In: *CVPR*. 2025.

[75] Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Sifan Long, Kun Kuang, Kun Yao, et al. “Hap: Structure-aware masked image modeling for human-centric perception”. In: *NeurIPS*. 2023.

[76] Xiao-Wen Zhang, Delong Zhang, Yi-Xing Peng, Zhi Ouyang, Jingke Meng, and Wei-Shi Zheng. “VIPerson: Flexibly Generating Virtual Identity for Person Re-Identification”. In: *ICCV*. 2025.

[77] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. “Scalable person re-identification: A benchmark”. In: *ICCV*. 2015.

[78] Tianyue Zheng and Weihong Deng. *Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments*. Tech. rep. Beijing University of Posts and Telecommunications, 2018.

[79] Tianyue Zheng, Weihong Deng, and Jiani Hu. “Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments”. In: *arXiv preprint arXiv:1708.08197* (2017).

[80] Jie Zhu, Xiao Guo, Yiyang Su, Anil Jain, and Xiaoming Liu. “FusionAgent: A Multimodal Agent with Dynamic Model Selection for Human Recognition”. In: *CVPR*. 2026.

[81] Jie Zhu, Yiyang Su, Minchul Kim, Anil Jain, and Xiaoming Liu. “A quality-guided mixture of score-fusion experts framework for human recognition”. In: *ICCV*. 2025.

[82] Jie Zhu, Ivy Zhang, Minchul Kim, and Xiaoming Liu. “Subtoken Vision Transformer for Fine-grained Recognition”. In: *arXiv preprint arXiv:2607.09086* (2026).

[83] Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. “Cross-covariate gait recognition: A benchmark”. In: *AAAI*. 2024.

[84] Jialong Zuo, Yongtai Deng, Mengdan Tan, Rui Jin, Dongyue Wu, Nong Sang, Liang Pan, and Changxin Gao. “ReID5o: Achieving Omni Multi-modal Person Re-identification in a Single Model”. In: *NeurIPS*. 2025.

A Additional Experiments

A.1 Additional Implementation Details

Architecture and K-SAH Specifications SapiensID 2.0 follows the SapiensID architecture: a 24-layer Vision Transformer (embedding dimension 512, 16 heads, 63.1M parameters) operating on 384×384 inputs with RetinaPatch, keypoint-based relative positional encoding over $K = 19$ keypoints, and a semantic attention head that outputs a 4096-d embedding. The full inference model has 411.5M parameters, of which 335.6M are in the linear layers that flatten the per-keypoint features into the final embedding; these layers are dense matrix multiplications and add little latency. The RP module explicitly isolates and normalizes semantic regions (e.g., face, full body, upper torso) before patch extraction, preserving absolute geometric coordinates via region-specific positional embeddings. The model is trained on static images only. K-SAH introduces no new parameters: at inference, the trained SAH weights are reused with a temporal window of $T = 3$ frames (stride 1) and mean pooling over the track, and the window size can be changed without retraining.

Semantic Subspace Distillation Setup The semantic feature spaces are generated entirely offline. We utilize the frozen `Qwen2.5-VL-7B-Instruct` [5] model to process the massive WebBody4M dataset. We employ the following structured prompt to ensure attribute consistency across instances:

```
"Analyze the individual in the provided image or video tracklet.
Generate a strict JSON object with:\n"
"group_a_fixed: race_ethnicity_presentation, gender_presentation,
age_demographic_category, "
"skin_tone, body_shape, facial_structure.\n"
"group_b_dynamic: hair_color_length_style, facial_hair_presence,
visible_accessories, "
"upper_clothing_color_type, lower_clothing_color_type.\n"
"If an attribute is obscured or ambiguous, return null for that key. "
"Do not infer uncertain information. Return JSON only."
```

The resulting text strings are passed through the `Qwen2.5-VL-7B-Instruct` [5] text encoder, producing high-dimensional textual embeddings. To obtain the requisite stable semantic topologies, we compute the covariance matrix over the entire WebBody training corpus and apply SVD independently to both the fixed and dynamic semantic sets. For ITA, we retain the top $D_F = 6$ principal components from the fixed biometrics subspace ($\mathbf{V}_T^F$). For TND, we extract the top $D_D = 5$ principal components from the dynamic biometrics subspace ($\mathbf{V}_T^D$). To align the visual embedding space with these lower-dimensional semantic bases, we attach a lightweight, two-layer Multi-Layer Perceptron (MLP) mapping the student feature dimension ($D \to 512 \to D$). This MLP serves exclusively to calculate the distillation gradients during training and is dropped prior to inference.

Training Hyperparameters The unified SapiensID 2.0 framework is optimized using the AdamW optimizer. We employ a base learning rate of 1×10^{-4}, accompanied by a cosine annealing learning rate schedule and a linear warmup over the first three epochs. Weight decay is set to 0.05. To effectively support the Relational Topology Preservation loss ($\mathcal{L}_{FR}$), we use a batch size of 104 per GPU on 7 NVIDIA H100 GPUs (total batch size 728). The model is trained for a total of 36 epochs. For the loss landscape, we use AdaFace [30], weighting our combined semantic distillation loss terms equally with a projection weight $\lambda_P = 0.5$ to ensure that the distillation process guides the latent space without overriding the fundamental metric learning constraints.

A.2 Datasets

In Tab. 5, we provide an overview of all the testing datasets that we use in our experiments. Note that we use one single model to test the zero-shot generalization capabilities of both SapiensID 2.0 and the baseline methods.

Table 5: Summary of testing datasets used in our experiments.

Dataset	Year	Task	Data Type	Same-Cloth.	Cloth.-Change	Other Variations
CCVID [17]	2022	ReID	Video	✓	✓	–
CCPG [36]	2023	ReID/Gait	Video		✓	Gait
LTCC [51]	2021	ReID	Image	✓	✓	–
PRCC [68]	2019	ReID	Image	✓	✓	–
Market1501	2015	ReID	Image	✓		–
MSMT17	2018	ReID	Image	✓		–
CCDA [40]	2023	ReID	Image		✓	–
Celeb-ReID [25]	2019	ReID	Image		✓	–
WebBody [32]	2025	ReID	Image			Pose/Scale
LFW	2007	Face	Image			–
CPLFW	2018	Face	Image			Pose
CFP-FP	2017	Face	Image			Pose
CALFW	2017	Face	Image			Age
AgeDB	2017	Face	Image			Age

Table 6: Generalization comparison with SoTA ReID models on two settings. "Long-term" refers to clothing change (CC) protocols of LTCC and PRCC datasets, while "short-term" refers to the same clothing (SC) protocols. For other datasets, the data capture characteristics define short or long-term conditions.

Method	Train Data	LTCC (General)		PRCC (SC) [68]		Market1501		MSMT17 [62]	
		top1	mAP	top1	mAP	top1	mAP	top1	mAP
CAL [17] (R50)	LTCC	74.04	40.84	99.51	95.64	35.60	16.11	15.92	5.06
CAL [17] (R50)	PRCC	20.69	6.19	**100.00**	99.76	18.97	6.47	2.56	0.69
CAL [17] (R50)	LTCC+PRCC	72.41	38.12	99.54	99.01	43.65	21.03	14.48	4.44
CLIP3DReID [41] (R50)	LTCC	75.66	**45.15**	99.43	96.43	41.66	20.33	17.45	5.50
CLIP3DReID [41] (R50)	PRCC	21.30	6.19	**100.00**	**99.84**	20.93	7.49	3.28	0.85
SOLIDER [8] (Swin-Base)	LUP4M+Market1501	73.83	36.28	99.51	**99.53**	**97.03**	**94.04**	48.64	22.77
SOLIDER [8] (Swin-Base)	LUP4M+MSMT17	74.44	36.74	99.30	98.71	89.85	73.20	**91.12**	**78.01**
HAP [75] (ViT-Base)	LUP4M+LTCC	65.11	29.02	95.53	86.44	51.63	27.29	20.89	6.56
HAP [75] (ViT-Base)	LUP4M+PRCC	63.29	29.36	98.84	98.38	73.49	50.11	29.61	10.99
HAP [75] (ViT-Base)	LUP4M+Market1501	73.02	35.97	99.30	98.45	96.23	92.20	48.01	23.02
HAP [75] (ViT-Base)	LUP4M+MSMT17	67.95	32.07	99.15	96.50	80.37	57.07	89.13	75.85
HAP [75] (ViT-Base)	WebBody	56.80	25.88	99.72	98.26	66.18	42.41	43.61	21.42
SapiensID [32] (ViT, 63M)	WebBody	72.01	34.56	**100.00**	98.79	88.18	68.26	67.25	31.02
SapiensID 2.0 (Ours, ViT, 63M)	WebBody	**77.08**	42.39	**100.00**	99.01	86.94	67.15	67.72	30.32

(a) Short-Term ReID

Method	Train Data	LTCC (CC) [51]		PRCC (CC)		CCDA [40]		Celeb-ReID [25]	
		top1	mAP	top1	mAP	top1	mAP	top1	mAP
CAL [17] (R50)	LTCC	38.01	18.84	37.00	35.20	3.91	9.67	37.42	3.92
CAL [17] (R50)	PRCC	6.38	3.14	55.69	55.64	2.85	8.61	23.59	2.20
CAL [17] (R50)	LTCC+PRCC	33.16	16.27	45.39	45.42	3.74	9.14	37.11	3.81
CLIP3DReID [41] (R50)	LTCC	41.84	22.58	40.81	38.38	4.31	10.18	37.31	4.02
CLIP3DReID [41] (R50)	PRCC	6.63	3.17	62.40	61.97	3.17	8.89	23.82	2.17
SOLIDER [8] (Swin-Base)	LUP4M+Market1501	25.00	12.18	26.87	32.12	8.62	16.48	46.37	5.66
SOLIDER [8] (Swin-Base)	LUP4M+MSMT17	26.02	11.33	22.27	25.36	8.79	15.54	47.95	6.14
HAP [75] (ViT-Base)	LUP4M+LTCC	25.00	11.63	26.14	22.34	4.56	11.18	30.28	3.54
HAP [75] (ViT-Base)	LUP4M+PRCC	29.08	12.52	38.05	41.94	5.13	13.40	37.79	4.48
HAP [75] (ViT-Base)	LUP4M+Market1501	24.74	11.71	33.90	37.00	8.30	16.02	44.38	5.20
HAP [75] (ViT-Base)	LUP4M+MSMT17	23.47	10.74	23.82	25.00	6.27	13.33	46.37	5.77
HAP [75] (ViT-Base)	WebBody	22.70	9.96	54.93	49.38	28.80	41.49	65.78	18.93
SapiensID [32] (ViT, 63M)	WebBody	42.35	17.79	78.75	72.60	61.84	69.08	92.80	66.92
SapiensID 2.0 (Ours, ViT, 63M)	WebBody	**50.26**	**24.28**	**87.19**	**81.32**	**66.97**	**73.06**	**93.64**	**73.87**

(b) Long-Term ReID

Table 7: ReID Performance on variable pose and scale settings on the WebBody test set.

Method	top1	mAP
CAL [17]	6.57	1.02
SOLIDER [8]	9.95	1.98
HAP [75]	8.22	1.52
SapiensID [32]	76.82	52.00
SapiensID 2.0 (Ours)	**78.44**	**55.03**

A.3 Additional Results

We show the results of additional baseline models with more diverse architectures trained on more diverse training datasets in Tab. 6. In the main paper, we report the results of the baseline model with the best overall generalization performance.

A.4 Performance of Gait Recognition Models on CCVID

To provide a comprehensive analysis of cross-domain capabilities, we evaluated recent large-scale gait foundation models, specifically BigGait [72] and BiggerGait [71], on the video-based person re-identification dataset (CCVID). We use models trained on CCGR [83], a large-scale gait recognition dataset commonly used for pretraining gait models with better generalization results than other gait datasets.

Surprisingly, gait recognition models demonstrate remarkably strong generalization to video-based ReID. BiggerGait achieves 85.13% Rank-1 and 83.38% mAP on the CCVID General split, vastly outperforming traditional ReID models like CAL (74.83% Rank-1) and CLIP3DReID (77.28% Rank-1). Most notably, in the clothing-change (CC) split, BiggerGait achieves a highly competitive mAP of 82.59%. This highlights the intrinsic advantage of gait architectures in appearance-altering scenarios: by relying entirely on kinematic silhouettes and discarding RGB texture, they are inherently immune to the clothing-change artifacts that catastrophically disrupt traditional ReID models like SOLIDER.

Despite this resilience, dedicated gait models are ultimately constrained by their modality limit. Because they discard all visual semantics (such as facial features, fixed soft biometrics, and unoccluded visual markers), they fail to match the peak discriminative capabilities of foundation models for human recognition. SapiensID 2.0 significantly outperforms BiggerGait in Top-1 accuracy across both the General (95.66% vs. 85.13%) and CC (88.63% vs. 84.53%) splits. This demonstrates that while BiggerGait is robust against noise, the optimal strategy for human recognition is a unified approach. By integrating kinematic continuity (via K-SAH) with explicit visual soft biometrics (via Invariant Trait Alignment), SapiensID 2.0 captures the temporal robustness of gait models without sacrificing the rich, discriminative power of static semantic identity.

A.5 Cross Pose-Scale ReID

To ensure that semantic alignment and kinematic tracking do not degrade the foundational spatial representations of the baseline, we evaluate the model on the challenging WebBody [32] test set, which features extreme variations in pose and scale. As shown in Table 7, traditional ReID models completely collapse (*e.g.*, SOLIDER yields only 9.95% Rank-1 accuracy). The SapiensID baseline exhibits strong pose and scale invariance due to its massive pre-training on WebBody4M. Crucially, SapiensID 2.0 maintains and further improves upon this robustness (78.44% Rank-1 and 55.03% mAP). This confirms that our Bipartite Subspace Alignment successfully disentangles semantic features without compromising the network's underlying geometric spatial priors.

A.6 Additional Ablation Studies and Benchmarks

Temporal Window Size Ablation in K-SAH To further dissect the impact of kinematic continuity, Tab. 8 illustrates the effect of the temporal window size (T) within K-SAH on the CCVID dataset (the $T \geq 3$ rows are also shown in Tab. 4 of the main paper). When $T = 1$, each frame attends only to itself and K-SAH reduces to the static SAH followed by temporal pooling. As the temporal window expands ($T \geq 3$), the network successfully tracks keypoint-guided semantic patches across neighboring

frames, capturing discriminative kinematic signatures that single-frame averaging completely discards. Performance saturates at $T = 3$, demonstrating that local temporal windows sufficiently capture kinematic continuity without incurring unnecessary computation.

Table 8: Impact of temporal tracking window size (T) within the K-SAH module on the CCVID dataset.

Window Size (T)	CCVID (General)		CCVID (CC)	
	top1	mAP	top1	mAP
SapiensID (frame mean)	92.57	77.82	88.72	72.22
3	95.66	85.00	88.63	78.05
5	95.65	84.95	88.52	77.95
7	95.70	84.90	88.50	77.87

Edge Latency and Inference Throughput Benchmark We benchmark SapiensID 2.0 against edge MLLMs on an NVIDIA H100 GPU (fp16, batch size 1). As shown in Tab. 9, edge MLLMs (MobileVLM V2 [10] and FastVLM [58]) suffer from low throughput due to autoregressive decoding of attribute tokens (> 5s latency per frame). In contrast, SapiensID 2.0 completely discards the teacher MLLM after training, running inference with a 411.5M-parameter network in a single forward pass with a latency of only 23.5ms (42.6 FPS), achieving a $224\times$ throughput speedup over MobileVLM V2.

Table 9: Inference latency, throughput, and parameter comparison with edge MLLMs (NVIDIA H100, fp16, batch size 1; SapiensID 2.0 averaged over 12 benchmark images). SapiensID 2.0 GFLOPs are per frame with K-SAH (T=3); GFLOPs and latency exclude the keypoint aligner (OpenPose and face detector, 72.3 GFLOPs).

Model	Parameters (Inference)	GFLOPs / Frame	Throughput (FPS)	Edge Latency
MobileVLM V2 (1.7B)	1.67B	3221.2	0.190	5.27s
FastVLM (1.5B)	1.68B	1967.3	0.162	6.17s
SapiensID 2.0 (Ours)	**411.5M**	**68.7**	**42.6**	**23.5ms**

Zero-Shot Out-of-Distribution Baseline Comparison To further evaluate out-of-distribution (OOD) clothes-changing generalization, Tab. 10 compares SapiensID 2.0 directly against specialized domain-specific zero-shot methods on the LTCC and PRCC clothing-change protocols. SapiensID 2.0 significantly outperforms recent zero-shot methods (e.g., 50.26% vs. 11.7% on LTCC, and 87.19% vs. 81.0% on PRCC), demonstrating the superior generalization of our decoupled semantic subspace alignment.

Table 10: Zero-shot out-of-distribution generalization comparison on LTCC (CC) and PRCC (CC). Baseline numbers are as reported in the original papers ("–": not reported); SapiensID 2.0 uses the same checkpoint and protocol as Tab. 2.

Method	Venue & Year	LTCC (CC)		PRCC (CC)	
		Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
CAL [17]	CVPR 2022	–	–	37.5	35.4
AIM [70]	CVPR 2023	–	–	40.7	38.4
GEFF [3]	WACV 2024	11.7	4.8	81.0	50.5
DLCR [52]	WACV 2025	–	–	45.5	38.8
SapiensID 2.0 (Ours)	**NeurIPS 2026**	**50.26**	**24.28**	**87.19**	**81.32**

A.7 Analysis of the Semantic Distillation

Examples of MLLM Soft-Biometric Annotations. Tab. 11 shows the soft-biometric descriptors generated by Qwen2.5-VL for the subject in Fig. 1 and for a randomly sampled WebBody training

image. The MLLM produces fine-grained descriptions of both the fixed traits used by ITA and the dynamic traits suppressed by TND, and returns `null` when an attribute is not visible, as instructed by the prompt.

Table 11: Examples of MLLM-generated soft-biometric annotations. Left: the subject in Fig. 1. Right: a random WebBody training image.

Group	Attribute	Fig. 1 subject	WebBody sample
Fixed	Race / ethnicity	White	Caucasian appearance
	Gender	Male	Female presentation
	Age	Adult	Young adult
	Skin tone	Fair	Light skin tone
	Body shape	Average	Medium build
	Facial structure	Average	`null`
Dynamic	Hair	Short, dark brown	Medium length dark brown hair
	Facial hair	`null`	`null`
	Accessories	`null`	`null`
	Upper clothing	Dark suit jacket	Blue long-sleeve sweater
	Lower clothing	Dark pants	Black trousers

Offline Probing Protocol. Retraining the full model for every distillation variant is expensive, so we additionally analyze the distillation targets and objectives with lightweight offline probes. We freeze SapiensID features (4096-d) on 10k WebBody images (8k for training, 2k for validation) and train a 2-layer MLP to map them to the semantic targets under different objectives. For these probes, the attribute strings are encoded with the CLIP ViT-L/14-336 text encoder [47]. We report the alignment error (MSE to the fixed soft-biometric targets) and the leakage of dynamic attributes into the learned features. These probes isolate the effect of the target and the objective, but they are not a substitute for end-to-end training.

Text Targets *vs*. Raw MLLM Hidden States. Instead of re-encoding the MLLM's textual output, one could distill directly from the MLLM's visual hidden states. We compare the two targets in Tab. 12. Leakage is measured with linear probes trained to predict clothing attributes from the features after projecting out the fixed-trait subspace: the residual clothing R^2 and the balanced accuracy of a 12-way upper-clothing color classifier (chance level 0.083). Raw Qwen2.5-VL visual hidden states (mean-pooled, 3584-d) already contain substantial clothing information (residual clothing R^2 of 0.217, compared to 0.003 for the text target $\mathbf{X}^F$), and distilling from them transfers this leakage to the student. Text targets act as a semantic filter: they retain only the queried attributes and discard background, pose, and other appearance cues, giving $4\times$ lower residual leakage at the same alignment error.

Subspace Alignment *vs*. Direct Regression. We compare ITA+TND against directly regressing the fixed-trait embeddings ($\|\mathbf{Y} - \mathbf{X}^F\|^2$) in Tab. 13. The orthogonality error is the energy of the projected features in the dynamic subspace $\mathbf{V}_T^D$. Both objectives reach the same alignment error, but direct regression leaves a substantial component along the dynamic (clothing) subspace, while ITA+TND removes it almost entirely. Since TND explicitly minimizes this quantity, the result mainly confirms that the dynamic subspace can be suppressed without sacrificing alignment with the fixed traits.

Table 12: Distillation target (offline probe).

Target	Align. MSE ↓	Leakage ↓	
		Res. R^2	Color
Raw MLLM hidden states	**0.045**	0.080	0.271
Text embeddings (Ours)	0.046	**0.020**	**0.179**

Table 13: Distillation objective (offline probe).

Objective	Align. MSE ↓	Orth. error ↓
Direct MSE	0.0441	0.0261
ITA + TND (Ours)	**0.0434**	$\mathbf{2.8 \times 10^{-6}}$

Identity *vs*. Clothing Clustering. To test whether the learned features cluster by identity rather than by outfit, we compute track-level cosine silhouette scores on CCVID (2,500 tracks), using either the identity or the clothing label as the cluster assignment. As shown in Tab. 14, SapiensID 2.0 increases

identity clustering and decreases clothing clustering, widening the identity–clothing gap from 0.004 to 0.040, well beyond the bootstrap standard error. Fig. 6 visualizes CCVID clip embeddings of SapiensID and SapiensID 2.0 for the same eight randomly chosen identities that appear in at least two outfits. For both models, clips form one cluster per identity, and different outfits of the same identity fall into the same cluster rather than into outfit-specific clusters; on these eight identities the two embeddings are qualitatively similar, and the difference between the models is the one measured over all tracks in Tab. 14.

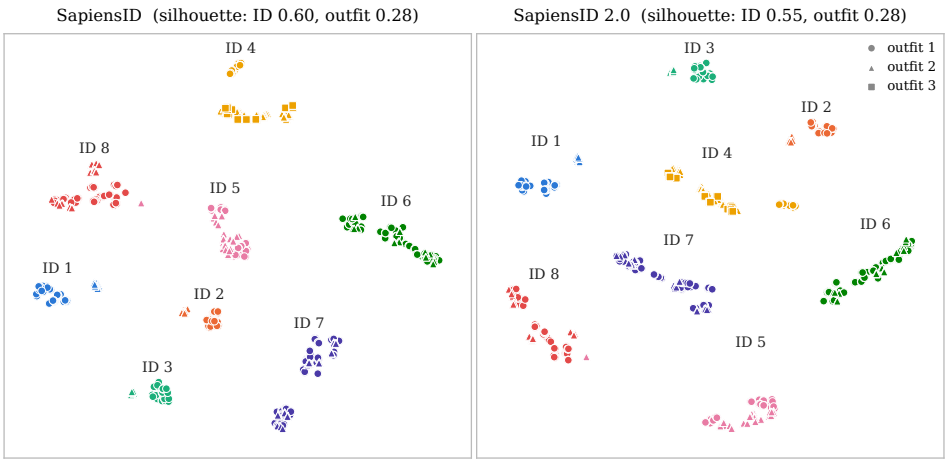


Figure 6: t-SNE of CCVID clip embeddings of SapiensID (left; mean over the 16 frames of a clip) and SapiensID 2.0 (right; K-SAH, T=3) for the same eight identities with at least two outfits. Colour denotes identity, marker shape the outfit. Panel titles give the cosine silhouette of these clips w.r.t. identity and outfit labels.

Where Does K-SAH Attend? We measure the mean K-SAH cross-attention weight per patch on stable landmarks (face, joints, body contour) and on clothing/background regions over 1,200 CCVID frames (Tab. 15). Both the static head and K-SAH concentrate attention on landmarks (over 2.2× the density on clothing/background), and extending the window to $T = 5$ preserves this spatial focus. The gain of K-SAH therefore comes from aggregating these landmark-anchored features across frames rather than from attending to different regions. Fig. 7 shows K-SAH cross-attention maps for one CCVID clip. Each frame of the window receives about one third of the attention mass, and the maps change little across the window. Face queries attend to a few face tokens, the shoulder query to the head and shoulder region, and limb queries to a horizontal band of body tokens at the height of the joint rather than to the joint alone.

Table 14: Track silhouette on CCVID (± bootstrap SE).

Model	Identity ↑	Clothing ↓
SapiensID	0.4438±.0037	0.4401±.0043
SapiensID 2.0	**0.4616**±.0038	**0.4219**±.0046

Table 15: K-SAH attention density on CCVID.

Head	Landmark	Cloth./bg	Ratio
Static SAH	1.450±.008	0.656	2.21
K-SAH (T=5)	1.447±.004	0.653	2.22

B Failure Cases and Trade-offs

Short-Term ReID. SapiensID 2.0 is slightly worse than SapiensID on several short-term metrics (Market1501 Rank-1 88.18% → 86.94%, MSMT17 mAP 31.02% → 30.32%) and on CCVID (CC) Rank-1 (88.72% → 88.63%). On short-term benchmarks the query and gallery share clothing, so clothing is a valid retrieval cue; TND deliberately suppresses it, which costs a small amount of accuracy in exchange for large gains under clothing change (Tab. 2b). Self-supervised models such as SOLIDER and HAP show the opposite trade-off: they are strongest on Market1501 but collapse under clothing change. Fig. 8 shows Market1501 queries that SapiensID retrieves correctly at rank 1



Figure 7: K-SAH cross-attention (T=3) for one CCVID clip. Columns: keypoint queries (green cross: the query keypoint, detected in the anchor frame t and drawn on every frame); rows: the frames of the window centred on frame t. The heatmap shows the positive part of the (single-head, differential) attention weight on each kept token (painted over the image region the token covers), averaged over the n queries per keypoint; the number in each panel is that frame's share of the attention mass over the window. Grey: padding outside the person crop.

while SapiensID 2.0 ranks the first correct match much lower (151 of 3,368 queries; the reverse holds for 54). In these cases the top-ranked images of SapiensID 2.0 are other people with similar clothing colour and build, or crops dominated by a bicycle, while SapiensID retrieves the same person.

MLLM Labeling Errors. The soft-biometric pseudo labels are not error-free. MLLMs can hallucinate attributes under occlusion or unusual poses, and when another person appears close to the subject they may attribute that person's hair or clothing to the subject. WebBody is collected with per-identity name queries, so most images contain a single dominant subject, which limits this contamination, but it does not eliminate it. Errors in fixed traits inject noise into the ITA targets, while errors in dynamic traits make TND suppress the wrong directions. Because both losses operate on the top principal components computed over the whole corpus rather than on individual labels, isolated errors have a limited effect on the subspaces, but systematic MLLM biases (*e.g.*, in perceived race or age) can propagate to the student.

C Potential Societal Impacts

The deployment of highly accurate, modality-unified human recognition systems carries profound societal implications. On the positive side, robust identity matching across extreme poses, occlusions, and clothing changes can drastically improve public security, aid in rigorous forensic investigations, and facilitate the recovery of missing persons in highly unconstrained environments. Conversely, the increased efficacy of these systems raises valid concerns regarding privacy infringement and the potential for unwarranted mass surveillance. Furthermore, because our semantic distillation relies on an MLLM to anchor traits such as race and gender, there is a distinct risk of inheriting and amplifying systemic biases present in the teacher model's pre-training data, which could lead to discriminatory profiling. It is imperative that the deployment of such foundational biometric models is accompanied by strict ethical guidelines, rigorous bias auditing, and transparent regulatory oversight to prevent misuse and ensure equitable societal outcomes.

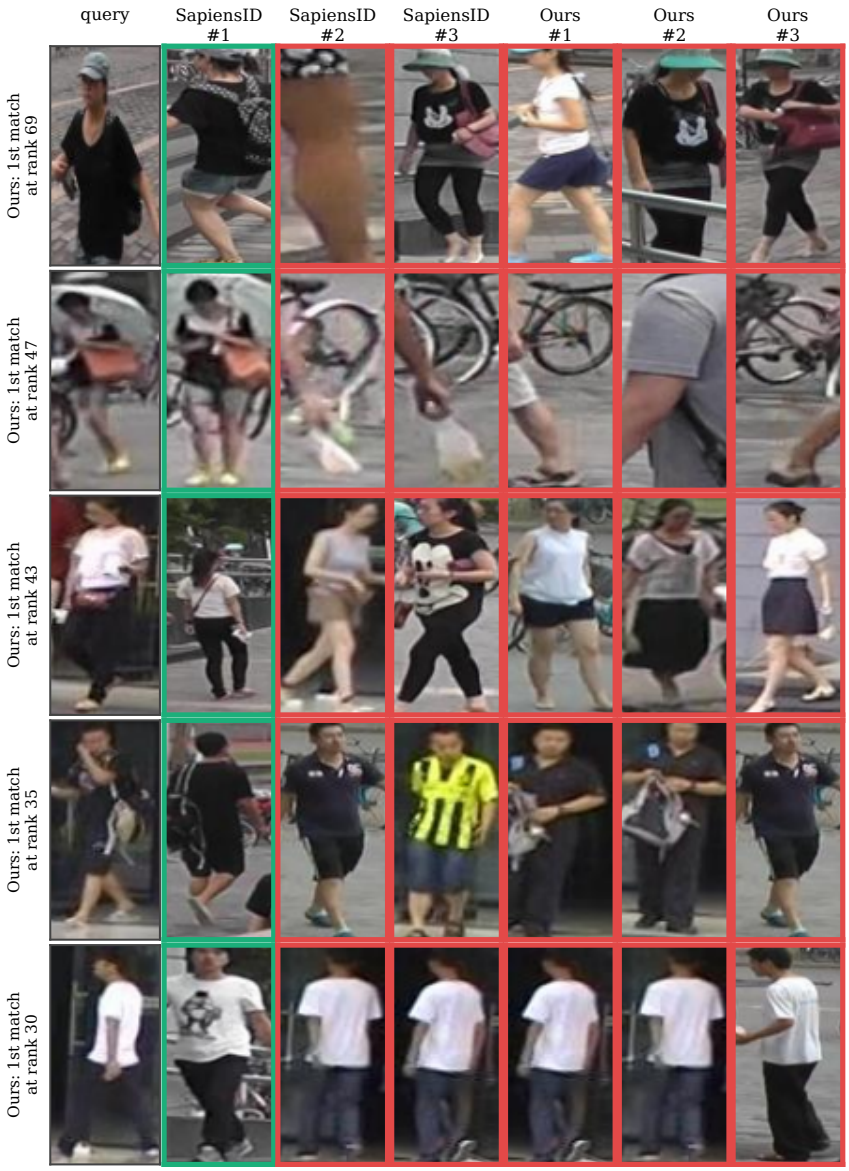


Figure 8: Market1501 failure cases of SapiensID 2.0 ("Ours"). Each row shows a query, the top-3 gallery images of SapiensID and of SapiensID 2.0 (green: same identity, red: different identity), and the rank of SapiensID 2.0's first correct match. Rows are the five such queries for which SapiensID 2.0 ranks the first correct match lowest.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline the paper's main contributions, including the proposed SapiensID 2.0 and its application to human recognition tasks. These claims are substantiated in the main body and experiments, reflecting both the scope and the actual contributions of the work.

   Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 5 (Limitations) discusses the limitations, with failure cases in the appendix (Failure Cases and Trade-offs): dependence on the teacher MLLM and its labeling errors, the offline extraction cost, and the small loss on short-term benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details regarding datasets, model architectures, training procedures, hyperparameters, and evaluation metrics. These details are adequate for reproducing the main experimental results that support the paper's conclusions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release inference code, the trained model weights, dataset preparation instructions and evaluation scripts that reproduce all reported SapiensID 2.0 results at `https://github.com/TODO/sapiensid2`. The evaluation datasets are obtained from their original sources. Training code and the WebBody training data are not released; WebBody is constructed as described in SapiensID [32].

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides sufficient details regarding datasets, model architectures, training procedures, hyperparameters, and evaluation metrics in Section 4 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Training is expensive (36 epochs on 7 H100 GPUs), so we train the final model once and do not report error bars over training runs. Evaluations use a fixed seed; for the clustering analysis in the appendix we report bootstrap standard errors.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix gives the GPU type and count (7 NVIDIA H100), batch size and number of epochs for training; the per-iteration training time and the inference latency are reported in Sec. 4 (Efficiency) and the appendix (Tab. 9).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only publicly available benchmarks under their terms of use and a training set constructed as in prior work; it involves biometric data, and the associated risks are discussed in the appendix (Potential Societal Impacts) and addressed by the release terms (see Safeguards).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The appendix (Potential Societal Impacts) discusses positive uses (e.g., forensics, finding missing persons) and risks (surveillance, privacy, bias inherited from the MLLM teacher).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the

technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The model weights are released for research use only under a licence that prohibits surveillance and other uses without a legal basis, with a model card describing intended use, limitations and known biases; no training images are redistributed.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets (datasets, codebases, and pretrained models) used in this work are properly cited. We have reviewed and complied with their respective licenses, and relevant details are included in the references or the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release an inference codebase and the trained model weights (`https://github.com/TODO/sapiensid2`) with a README (installation, data preparation, evaluation commands, expected results), a model card and a licence.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on existing datasets and benchmarks, and no new human participants are recruited for data collection, annotation, evaluation, or user studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve any human subjects or newly collected human data, and therefore does not require IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Qwen2.5-VL-7B-Instruct is used offline to annotate soft biometrics of the training images, which define the distillation targets (Sec. 3 and appendix).

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.