BEYOND FRAME-WISE ENCODING: WHEN DO VIDEO-NATIVE REPRESENTATIONS HELP GAIT RECOGNITION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision models (LVMs) have advanced gait recognition, but most pipelines retain the GaitSet-style design: an image LVM encodes each frame independently before a gait head aggregates the sequence. Video LVMs instead encode clips jointly, producing *video-native representations* shaped by cross-frame context. We study when these representations help and what they change before the gait head. Across several benchmarks, frozen video LVMs remain competitive on within-domain evaluations and transfer better to unseen domains. To isolate the effect of clip-wise encoding, we evaluate the same V-JEPA 2.1 checkpoint frame-wise and clip-wise, using the same training samples and gait-head architecture. Clip-wise encoding has little effect on within-domain accuracy but improves cross-domain recognition, with the largest gains under nighttime capture and clothing changes. Under controlled occlusion, its advantage depends on temporal coherence rather than on whether hidden regions reappear: the advantage grows when a body region remains hidden throughout the sequence but reverses when visibility shifts randomly across frames, in which case frame-wise representations perform better. At the frozen-encoder level, clip-wise context stabilizes representations of still-visible regions and makes unchanged observations more identity-discriminative, thereby changing what the gait head pools rather than merely how it pools. Gait research has long focused on refining how features are pooled; video-native representations offer a way to improve the features themselves. Our results identify where this benefit emerges and where clip-wise context becomes a liability, positioning video-native representations as a promising foundation for robust gait recognition.

1 INTRODUCTION

Gait recognition identifies individuals by their unique walking patterns. Unlike face, fingerprint, or iris recognition, it works at a distance and without the subject's cooperation (Shen et al., 2025). Its core challenge is that the *identity evidence* in a walk, the cues that distinguish one person from another, must survive covariates such as clothing change, carrying, viewpoint, and occlusion. Early methods convert frames into silhouettes or skeletons (Chao et al., 2019; Fan et al., 2024b), which suppress clothing but also discard identity cues (Ye et al., 2025); recent RGB methods instead build on frozen large vision models (LVMs), which preserve more of the person and have improved both within-domain and cross-domain recognition (Ye et al., 2024; 2025; Huang et al., 2026a).

These LVM pipelines inherit the view of gait as a set (Chao et al., 2019): frames are encoded independently and pooled without regard to order, a view that remains influential and has been explicitly defended (Hou et al., 2025). Where gait models do capture temporal structure, it is learned by gait-specific architectures, through part-based temporal modules, 3D convolutions, snippets, or transformer backbones (Fan et al., 2020; 2024a; Hou et al., 2026b;a), or added on top of frame-wise LVM features as a motion branch (Huang et al., 2026a). None of them builds on a representation whose cross-frame context is learned from large-scale video: in LVM pipelines, the features of each frame are fixed before the rest of the walk is seen (Fig. 1a).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

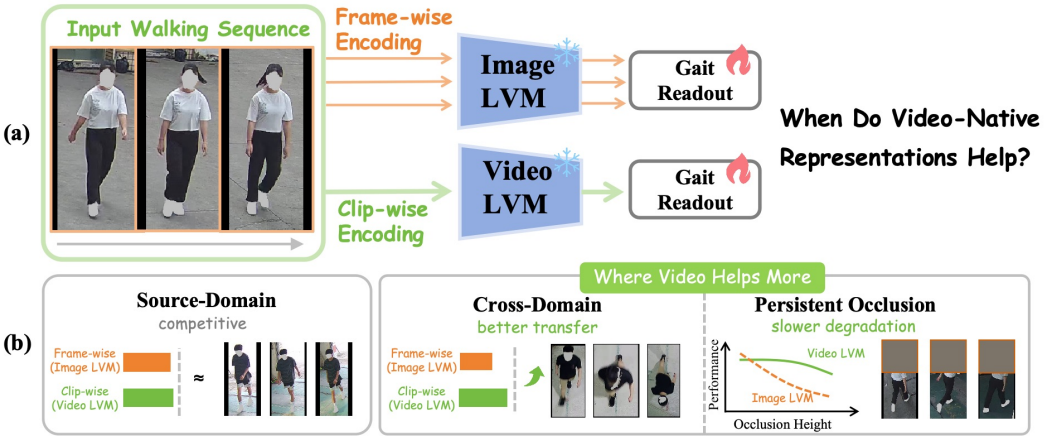

Figure 1: **Frame-wise versus clip-wise foundation representations.** (a) Image LVMs encode frames independently before a common gait head; video LVMs encode the clip jointly, moving cross-frame interaction into the frozen representation. (b) Video-native representations remain competitive in the source domain, while transferring better to unseen domains and degrading more slowly under persistent occlusion. These findings suggest that clip-wise encoding is particularly beneficial when identity evidence is unreliable.

Video LVMs such as V-JEPA 2.1 (Mur-Labadia et al., 2026) and InternVideo-Next (Wang et al., 2025a) provide such a representation. They encode a clip jointly, so context shapes the features of each frame before pooling; we call this *clip-wise encoding* and its outputs *video-native representations*. Whether this helps gait is not obvious: video pretraining targets generic scenes rather than walking people, and joint encoding may entangle identity cues with nuisance motion. We therefore ask: *When do video-native representations help gait recognition, what do they change, and what does this imply for future gait algorithms?*

We look for the answer where identity evidence is unreliable, since clean within-domain rankings reveal little about whether a model can compensate for missing evidence. Unseen domains change appearance. Occlusion removes part of the body: it is a long-standing difficulty for gait recognition and person re-identification (Xu et al., 2023b;a; Huang et al., 2024; Gupta & Chellappa, 2024; Hou et al., 2019), and, unlike domain shift, it can be controlled exactly, deciding which body part is hidden and for how long.

We first evaluate frozen video LVMs on four gait datasets under within-domain and cross-domain protocols. V-JEPA-L achieves the best cross-domain Rank-1 on all three unseen datasets, but such comparisons confound the encoding with the backbone and its pretraining. To isolate the encoding itself, we run one V-JEPA 2.1 checkpoint, pretrained on both images and videos, in its two native modes, frame-wise through its image pathway and clip-wise through its video pathway, with everything else matched. Measured against the set view, three findings emerge.

Temporal context helps most where appearance fails. The set view assumes that the appearance of a frame already implies its place in the gait cycle, so frames can be encoded independently (Chao et al., 2019). Within-domain, this holds: frame-wise and clip-wise encoding of the same checkpoint differ by only +1.43% Rank-1 on CCPG CL. Across domains they diverge, most where appearance is least reliable: clip-wise encoding gains +11.41% under night-time capture on SUSTech1K and +4.45% under clothing change on CASIA-B*, the hardest condition for frame-wise features in each dataset, and nothing on CASIA-B* NM and BG, where clothing is unchanged (§3).

Temporal context needs coherent visibility, not re-exposure. Under the set view, a walk resists occlusion through re-exposure: set-style pooling collects each body part from whichever frames show it. Clip-wise encoding behaves the opposite way. Its advantage is largest when a part never reappears, reaching +7.9% Rank-1 at 60% occlusion averaged over body locations, and reverses when hidden parts reappear at random, the condition with the most re-exposure; there, frame-wise encoding, which cannot suffer from incoherence, is the better choice (§4). What matters is whether visibility evolves coherently, not how often hidden parts return.

Temporal context reshapes what the gait head pools. Under the set view, frame features are fixed elements, and progress comes from better pooling. Clip-wise encoding changes the elements themselves. Before any pooling, its features drift 16.4% less than frame-wise ones where the body stays visible yet change more where it is hidden, so context protects what remains rather than filling in what is missing. A pair of frames held bitwise fixed also becomes more identity-discriminative inside its natural walk than inside a shuffled one; since the gait head sees only that pair, the gain cannot come from pooling (§5).

Together, these findings qualify rather than overturn the set view: set-level pooling can stay, but its elements are better encoded in context, as long as that context is coherent. For gait research, they suggest that temporal modelling may be better placed in the representation than in the gait head, and that video LVMs are a natural foundation for gait recognition where identity evidence is unreliable (§6).

2 Study Design

2.1 Frame-wise and Clip-wise Encoding under a Common Gait Head

To isolate the frozen foundation representation, we keep the gait head of BiggerGait (Ye et al., 2025) for every model. For a sequence X and encoding mode $e \in \{\text{frame}, \text{clip}\}$, the embedding is

$$z^e(X) = \operatorname*{cat}_{\ell \in \mathcal{L}} G^e_\ell\big(A^e_\ell(E^e_\ell(X))\big), \tag{1}$$

where E^e_ℓ is the frozen representation from encoder layer ℓ, A^e_ℓ maps it to a human-centred space with a learned foreground gate, and G^e_ℓ is a GaitBase network (Fan et al., 2023) (ResNet-9, temporal max pooling, 32-bin horizontal pyramid pooling, part-specific fully connected layers, and a BN neck). Only A^e_ℓ and G^e_ℓ are trained with equal-weight triplet and cross-entropy losses; modes share their architecture and training recipe but use separate parameters. The *frame-wise* mode encodes each frame independently; the *clip-wise* mode encodes an ordered 32-frame clip jointly. We call the features of one time step *frame-level features*; for clip-wise encoding, a time step is a two-frame tubelet. The gait head max-pools these features over time without order, retaining the set view so temporal-context effects must arise in the encoder. In the matched comparison, both modes use the same contiguous 32-frame training samples and differ only in encoding.

Representations. V-JEPA 2.1-B/L (Mur-Labadia et al., 2026) was pretrained jointly on images and videos, with a 2D patch embedding and an image modality token for images and 3D two-frame tubelets for video. Its frame-wise mode passes every frame through the image pathway and its clip-wise mode passes the clip through the video pathway, so both are native uses of the same weights; this makes it suited to a matched comparison, which earlier work obtained by training a separate image model (Vélez et al., 2025). The two pathways necessarily tokenize differently (a patch per frame versus a tubelet per two frames). The organization-sensitivity and fixed-observation interventions (Sections 5.1 and 5.3) test contextual effects within the video pathway; the latter holds the anchor's pixels, position, and input tokenization fixed. InternVideo-Next-B (Wang et al., 2025a) is a video LVM run clip-wise, and DINOv2-S/L (Oquab et al., 2023) are image LVMs run frame-wise. They place video LVMs on the benchmark and add a cross-family check under persistent occlusion, for which DINOv2-S is our own run under the same pipeline (clean CL 89.5%). We report robustness as absolute Rank-1, as *retention* (Rank-1 relative to the same mode's clean score), or as the *clip-wise advantage* (clip-wise minus frame-wise Rank-1), and say which.

2.2 Data, Protocols, and Metrics

Following BiggerGait, which chose CCPG (Li et al., 2023) for its rich clothing variation, we train on its 100 training identities and test on its 100 test identities under full clothing change (CL), top change (UP), pants change (DN), and bag change (BG). Without target-domain training, we evaluate on three datasets that vary appearance differently: CCGR-MINI (Zou et al., 2024) spans many covariates, the RGB modality of SUSTech1K (Shen et al., 2023) has night-time and uniform subsets, and CASIA-B* (Yu et al., 2006) separates clothing change from normal walking and bag carrying, all under the OpenGait/BiggerGait gallery–probe protocols (Fan et al., 2023; Ye et al., 2025). Unless noted, we report Rank-1 (%) excluding identical-view matches and write gaps as absolute Rank-1

Table 1: **Positioning on the benchmark: source and cross-domain Rank-1 (%).** All methods are trained on CCPG (Li et al., 2023); CCPG is evaluated within-domain, whereas CCGR-MINI (Zou et al., 2024), SUSTech1K (Shen et al., 2023), and CASIA-B* (Yu et al., 2006) are evaluated without target-domain fine-tuning. Published methods are from (Chao et al., 2019; Lin et al., 2022; Fan et al., 2024a;b; Zheng et al., 2024; Jin et al., 2025a; Liang et al., 2022; Jin et al., 2025b; Wang et al., 2025b; Ye et al., 2024; 2025; Huang et al., 2026a). Rows above the midrule are published results; rows below are our runs of frozen video LVMs (Mur-Labadia et al., 2026; Wang et al., 2025a) with the BiggerGait gait head (Ye et al., 2025), used as analysis probes rather than a proposed method. Rows differ in backbone, pretraining data, size, and feature design, so this table positions pipelines and does not attribute their differences; Table 2 provides the checkpoint-matched comparison. Avg. follows BiggerGait: the mean of the CCPG average, CCGR-MINI R1, SUSTech1K R1, and the CASIA-B* average. Best and second-best in **bold** and underlined; orange and blue cells mark the within-domain and cross-domain best.

Method	Venue	Input	LVM	Source domain				Cross-domain targets							Avg.
				CCPG				CCGR-MINI		SUSTech1K		CASIA-B*			
				CL	UP	DN	BG	R1	R5	CL	R1	NM	BG	CL	
GaitSet	AAAI'19	Sil.	–	60.2	65.2	65.1	68.5	2.4	6.9	8.2	12.8	47.4	40.9	25.8	29.5
GaitGL	arXiv'22	Sil.	–	68.3	76.2	67.0	76.7	3.3	8.4	25.4	33.6	63.1	58.5	46.3	41.2
DeepGaitV2	arXiv'23	Sil.	–	78.6	84.8	80.7	89.2	3.7	9.1	27.0	38.4	74.6	67.2	50.2	47.4
SkeletonGait++	AAAI'24	Sil.+Skel.	–	79.1	83.9	81.7	89.9	–	–	–	–	–	–	–	–
XGait	MM'24	Sil.+Par.	–	72.8	77.0	79.1	80.5	–	–	–	–	–	–	–	–
MultiGait++	AAAI'25	Sil.+R.+Flow	–	83.9	86.8	86.7	91.5	–	–	–	–	–	–	–	–
GaitEdge	ECCV'22	RGB+Sil.	–	66.9	74.0	70.6	77.1	–	–	8.9	19.6	66.5	58.7	44.8	–
DenoisingGait	CVPR'25	RGB+Sil.	SD 1.5	84.0	88.0	90.1	95.9	–	–	37.3	59.1	83.9	76.1	34.8	–
Gait-X	ICCV'25	RGB	–	82.0	86.0	90.7	95.7	–	–	47.9	73.4	81.8	76.9	28.5	–
BigGait	CVPR'24	RGB	DINOv2-S	82.6	85.9	87.1	93.1	7.4	16.3	43.7	56.4	77.4	71.5	33.6	53.0
BiggerGait	NeurIPS'25	RGB	DINOv2-S	89.8	92.1	93.7	97.5	15.5	26.9	70.9	79.6	93.0	90.8	55.6	67.0
BiggerGait	NeurIPS'25	RGB	DINOv2-L	**92.8**	**94.6**	**95.7**	**98.2**	17.7	26.9	76.4	85.3	95.4	93.7	62.2	70.5
GaitMax	CVPR'26	RGB	DINOv3-S	86.6	88.2	90.2	93.2	–	–	55.0	–	85.6	86.9	46.2	–
Ours	–	RGB	InternVideo-Next-B	83.0	87.4	90.5	97.2	22.7	**35.6**	70.5	87.7	94.9	91.7	49.7	69.7
Ours	–	RGB	V-JEPA-B (clip.)	86.8	91.0	93.4	98.1	19.4	31.3	75.7	90.9	96.8	94.6	52.1	70.9
Ours	–	RGB	V-JEPA-L (clip.)	88.4	92.5	94.1	98.1	**23.0**	**35.6**	**82.3**	**94.6**	**98.1**	**97.2**	**62.9**	**74.2**

differences (e.g. +1.43%); CCPG CL is the primary metric, and in CCPG diagnostics "Avg" averages CL, UP, DN, and BG. All encoders are frozen and the gait-head schedule is identical across models; resolutions, sampling, optimization, and compute are in Appendix A.

3 Temporal Context Helps Most Where Appearance Fails

The set view assumes that the appearance of a frame already implies its place in the gait cycle (Chao et al., 2019). If so, encoding frames in context should add little; if context supplies identity evidence that a single frame lacks, it should add most where appearance is unreliable. Table 1 first positions frozen video LVMs; a matched comparison then tests this prediction.

3.1 Positioning: Video LVMs Transfer Better Across Domains

Table 1 places frozen video LVMs, run with the BiggerGait gait head, among published methods trained on CCPG. Within-domain, they are competitive but not best: V-JEPA-L reaches 88.4% on CCPG CL, below BiggerGait with DINOv2-S (89.8%) and DINOv2-L (92.8%). Across domains, V-JEPA-L has the highest Rank-1 on CCGR-MINI (23.0% vs. 17.7% for the best image LVM) and SUSTech1K (94.6% vs. 85.3%), and leads on CASIA-B* (62.9% vs. 62.2% on CL, with larger margins on NM and BG), with fewer parameters and GFLOPs than DINOv2-L (Appendix Table 4). The profile is not universal: InternVideo-Next exceeds BiggerGait with DINOv2-S on CCGR-MINI and on SUSTech1K overall, but not on CASIA-B* CL. These rows also differ in pretraining, capacity, and feature selection, so the table motivates, rather than replaces, a matched comparison.

3.2 Attribution: One Checkpoint, Two Encoding Modes

Table 2 runs one V-JEPA 2.1-B checkpoint in both modes. Clip-wise encoding barely changes within-domain accuracy: +1.43% on CCPG CL, and +1.23%, +0.19%, and +0.21% on UP, DN,

Table 2: **Attribution: matched V-JEPA 2.1-B encoding modes** (Rank-1 %). Both modes share weights, layers, resolution, contiguous 32-frame training samples, and gait head; frame-wise encoding uses the native image pathway and clip-wise encoding the video pathway. Gap = clip-wise − frame-wise, computed before rounding.

Encoding	CCPG CL	CCGR R1	SUSTech R1	CASIA-B* CL
Frame-wise	85.35	17.03	87.76	47.66
Clip-wise	86.78	19.36	90.87	52.11
Gap	+1.43	+2.33	+3.11	+4.45

and BG. The within-domain ordering in Table 1 is therefore a difference between model families, not a cost of clip-wise encoding. Across domains, clip-wise encoding is higher on 12 of 14 metrics, including CCGR-MINI R1 (+2.33%), SUSTech1K overall (+3.11%), and CASIA-B* CL (+4.45%); the two exceptions, CASIA-B* NM and BG, differ by less than 0.1%.

Where the gain concentrates is informative. In each dataset with sub-conditions, it is largest on the condition where frame-wise features are weakest: night-time capture on SUSTech1K (+11.41%), clothing change on CASIA-B* (+4.45%, against −0.08% and −0.06% on NM and BG, where probe and gallery wear the same clothes), and, by a small margin, CL on CCPG. These are also the conditions in which appearance says least about who is walking, as low light degrades colour and texture and changed clothing breaks the appearance match; uniform clothing on SUSTech1K (+6.01%) points the same way. This is where the set view's assumption breaks: a single frame carries no motion, so frames encoded in isolation must rely on appearance and shape. A within-domain ranking, where all gaps stay below 1.5%, would call the two encodings equivalent.

Finding 1. Temporal context helps most where appearance fails. On the same checkpoint, clip-wise encoding gains under 1.5% Rank-1 within-domain, but +11.41% under night-time capture and +4.45% under clothing change, the conditions in which a single frame says least about who is walking. A within-domain ranking would call the two encodings equivalent.

4 Temporal Context Needs Coherent Visibility, Not Re-exposure

Occlusion lets us decide which body part is hidden and for how long, and thereby separates two explanations of why video might help. The set view suggests *re-exposure*: a part hidden in one frame can be collected from another, as set-style pooling does and as occluded video re-identification does when it completes hidden regions from other frames (Hou et al., 2019; 2022). The alternative is *coherence*: context from neighbouring frames strengthens the evidence that remains visible, which requires visibility to evolve coherently over time.

Three occlusion patterns separate the two explanations. All patterns are horizontal bands applied to probes only, as in earlier occluded-gait benchmarks (Uddin et al., 2019; Gupta & Chellappa, 2024) (Fig. 2a). *Persistent* bands hide the same region in every frame, as static occluders and the frame border do: no re-exposure, coherent visibility. *Smooth* bands move gradually, as occluders whose coverage changes during the walk: re-exposure with coherent visibility. *Random* bands are placed independently in every frame: the most re-exposure, but incoherent visibility. Random bands have no physical counterpart and serve as a diagnostic, following the artificial random occlusion of Uddin et al. (2019). If the clip-wise advantage came from re-exposure, it should be smallest under Persistent bands and largest under Random ones; if it comes from coherence, the order reverses. Top, middle, and bottom bands hide roughly the head and shoulders, the torso and arms, and the legs; masks are filled with the ImageNet mean and identical for all models (Appendix B).

Persistent bands: the largest gain, with no re-exposure. Clip-wise V-JEPA outperforms frame-wise V-JEPA in all nine location–severity conditions and on every CCPG protocol (Appendix Fig. 4). Averaged over locations, its CL lead grows from +1.4% on clean probes to +6.4%, +7.3%, and +7.9% at 20%, 40%, and 60% of body height (37.2% vs. 29.3% at 60%; Fig. 2b). The hidden band never returns, so this advantage cannot come from recovering it from other frames. InternVideo-

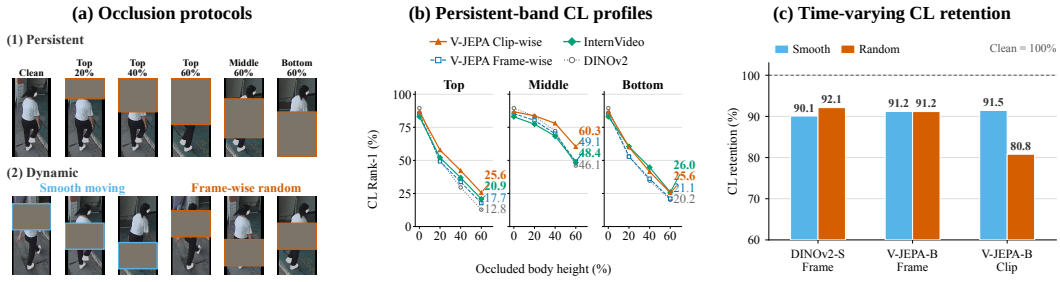

Figure 2: **Controlled occlusion on CCPG probes; the gallery stays clean.** (a) Persistent, Smooth, and Random bands differ in whether hidden regions reappear and whether visibility evolves coherently. (b) Persistent bands covering 20–60% of body height: CL Rank-1 for the matched V-JEPA 2.1-B modes and the cross-family InternVideo-Next/DINOv2-S comparison; shallower declines are more robust. (c) Time-varying 40% bands: CL retention relative to each mode's clean score (absolute values in Appendix Table 5).

Next, a video LVM with a different pretraining recipe, also loses fewer CL points than DINOv2-S in all nine conditions.

Moving bands: coherent return shrinks the gain, random return reverses it. Smooth and Random bands both hide 40% of body height in every frame. Under Smooth bands, the clip-wise advantage shrinks from +7.3% under Persistent bands of the same size to +1.6% (79.4% vs. 77.8% CL), close to its clean-probe level, and both modes retain about 91% of their clean score. Under Random bands, frame-wise V-JEPA is unchanged (77.8%) and DINOv2-S even improves (80.6%→82.4%), since set-style pooling collects each region from whichever frames show it; clip-wise V-JEPA instead drops to 70.1%, or 80.8% of its clean score, and falls 7.7% behind (Fig. 2c; Appendix Table 5). The clip-wise advantage is therefore largest under Persistent bands, smaller under Smooth ones, and reversed under Random ones, the order that coherence predicts and re-exposure does not. Why does incoherent visibility hurt? The organization controls in Section 5.1 show that clip-wise recognition changes little under chunk reversal but degrades under frame shuffling, whereas frame-wise recognition is invariant to full shuffling. V-JEPA 2.1 is pretrained to predict masked spatiotemporal content (Mur-Labadia et al., 2026), so its encoder draws on neighbouring frames that it expects to be consistent. Randomly jumping bands likely carry the corruption of each frame into the features of its neighbours; part of the penalty may also arise at the input, where each two-frame tubelet mixes frames with differently placed bands. Independently encoded frames are immune by construction, which is where the set view pays off. The patterns also differ in local transitions and visibility durations, so this identifies a failure condition rather than isolating one mechanism.

Finding 2. Temporal context needs coherent visibility, not re-exposure. The clip-wise advantage is largest when a hidden body part never reappears (+7.9% CL Rank-1 at 60% occlusion) and reverses when hidden parts reappear at random, where independently encoded frames are the safer choice. What matters is how visibility evolves over time, not how often hidden parts return.

5 Temporal Context Reshapes What the Gait Head Pools

Under the set view, frame features are fixed elements and only their pooling can be improved. We test this assumption in three steps: whether clip-wise features depend on the surrounding frames at all, what that dependence does to the frozen tokens when part of the body is hidden, and whether it makes an unchanged frame more useful for identity.

5.1 Clip-wise Features Depend on Frame Order

GaitSet showed that gait can be recognized from an unordered set of frames (Chao et al., 2019); frame shuffling (Misra et al., 2016) asks whether clip-wise features also treat a walk as a set. On

Table 3: **Sensitivity to clip organization** for the matched V-JEPA 2.1-B checkpoint on CCPG (Rank-1 %). Probes are reordered at test time; the gallery stays clean. Drops are computed before rounding.

		Rank-1 (%)		Drop from clean	
Encoding	Probe order	CL	Avg.	CL	Avg.
Clip-wise	Clean	86.8	92.3	0.0	0.0
Clip-wise	Chunk reverse	86.2	91.9	0.6	0.5
Clip-wise	Chunk shuffle	64.8	78.3	21.9	14.0
Clip-wise	Full shuffle	48.2	65.6	38.6	26.7
Frame-wise	Clean	85.4	91.6	0.0	0.0
Frame-wise	Full shuffle	85.4	91.6	0.0	0.0

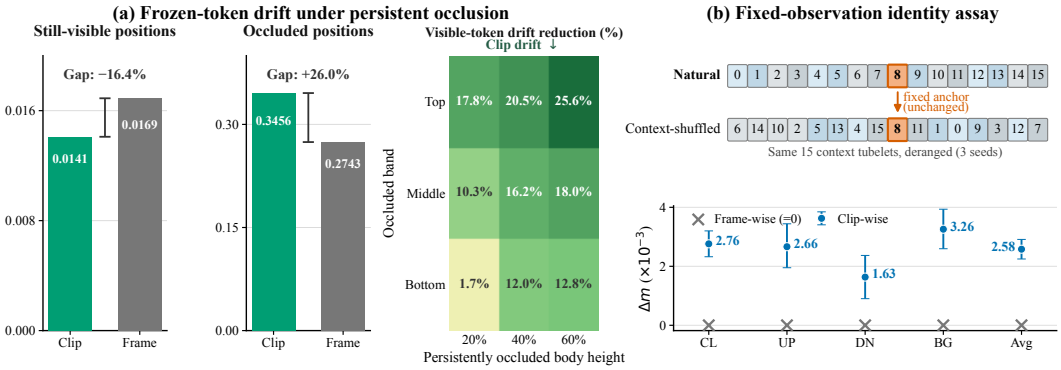

Figure 3: **What clip-wise encoding changes in frozen V-JEPA 2.1-B tokens.** (a) Mean cosine drift (Eq. 2) at still-visible and overwritten positions, averaged over four layers and nine persistent masks, and the relative reduction of visible-position drift, $(d_{\text{frame}} - d_{\text{clip}})/d_{\text{frame}}$, for each mask. (b) Tubelet 8 is kept bitwise identical and in place while the other 15 tubelets are deranged (one of three derangements shown); the gait head receives only the anchor. Δm (Eq. 3) with identity-bootstrap 95% intervals; frame-wise encoding is exactly invariant.

clean probes, we reverse each 32-frame chunk, permute frames within chunks, or permute them across the whole sequence; the gallery stays clean (Table 3).

Frame-wise V-JEPA is invariant by construction, because it encodes frames independently before an order-invariant gait head: full shuffling leaves CL and Avg. unchanged at 85.4 and 91.6. Clip-wise V-JEPA changes little under chunk reversal (-0.6 CL) but loses 21.9 CL points under chunk shuffle and 38.6 under full shuffle. Its features therefore depend on local temporal organization while remaining largely insensitive to its direction; they are not a bag of independently encoded frames. Because shuffled clips are out of distribution, these controls establish sensitivity to clip organization rather than identifying a specific motion primitive or independently explaining the occlusion-robustness gap.

5.2 Context Protects What Stays Visible

If clip-wise encoding filled in hidden content from other frames, its tokens would drift less where pixels were overwritten; if it protected visible evidence, where the body stays visible. For encoding mode e, layer ℓ, and native token coordinate u, we measure

$$d_{\cos}^{e,\ell}(u) = 1 - \cos\left(E_\ell^e(X)_u,\ E_\ell^e(\widetilde{X})_u\right), \tag{2}$$

where $\widetilde{X}$ is the persistently occluded version of the same 32-frame sample. Distances are taken before any trained component. Coordinates are labelled still-visible or overwritten by patch coverage, and frame-wise distances are aligned to the two-frame tubelet support after measurement. The analysis covers 2,329 paired sequences, all 100 test identities, nine masks, and four layers ($\{2, 5, 8, 11\}$).

At still-visible positions, clip-wise tokens drift 16.4% less than frame-wise tokens (cosine drift 0.0141 vs. 0.0169; the paired identity-bootstrap 95% interval of the difference excludes zero, Appendix Table 6). The direction holds for all nine masks and all 100 identities, and relative-ℓ_2 distance gives the same result (Fig. 3a). At overwritten positions the ordering reverses: clip-wise tokens change 26.0% more (0.346 vs. 0.274). The evidence thus favours protection over filling-in: clip-wise encoding reduces change where the body is still visible and stays responsive where the input changed.

5.3 Context Makes the Same Frames More Discriminative

Token stability alone does not show that the preserved information is useful for identity. We therefore hold a local observation fixed and ask whether its surrounding clip changes its identity value. Each 32-frame clip contains 16 two-frame tubelets. We keep tubelet 8 (frames 16–17) bitwise identical and at the same absolute position, and permute only the other 15 tubelets with three fixed derangements (Fig. 3b). The anchor's tokenization is therefore unchanged, and only its temporal context differs. We take the anchor's tokens from the same four layers and pass only the anchor, with $T{=}1$, through the trained gait head, which never sees the surrounding tubelets and serves as a fixed identity assay. With s^+ the mean similarity of the anchor embedding to the gallery sequences of the correct identity and s^- the largest such mean over the other identities ,

$$m = s^+ - s^-, \qquad \Delta m = m_{\text{natural}} - \tfrac{1}{3}\textstyle\sum_{k=1}^{3} m_{\text{deranged},k}. \tag{3}$$

Positive Δm means that natural context makes the identical anchor more discriminative. Effects are averaged within each test identity before 10,000 identity-level bootstrap resamples.

For clip-wise V-JEPA, Δm is positive for every CCPG protocol: 2.76×10^{-3} on CL (95% CI $[2.32, 3.20] \times 10^{-3}$) and 2.58×10^{-3} on Avg ($[2.25, 2.91] \times 10^{-3}$); frame-wise encoding is exactly zero. Because the anchor's pixels, its position, and the assay are fixed, the identity-relevant change originates in the frozen, context-conditioned features; it cannot be explained by the gait head pooling additional frames. Clip-wise encoding thus changes what the gait head pools, not only how it pools. Under persistent occlusion, this context benefit grows when the legs stay visible and vanishes when they are hidden (Appendix C.3).

Finding 3. Temporal context reshapes what the gait head pools. Before any pooling, clip-wise tokens drift 16.4% less where the body stays visible and 26.0% more where it is hidden, and a pair of frames held bitwise fixed becomes more identity-discriminative inside its natural walk than inside a shuffled one. Context improves the elements of the set, not only their pooling.

6 Discussion

Why context helps, and when it hurts. The controls in Appendix D leave one account for all three findings: the temporal prior of video pretraining. V-JEPA 2.1 learns by predicting masked spatiotemporal content (Mur-Labadia et al., 2026), so its encoder draws on neighbouring frames when a single frame is uninformative. This explains the gains where appearance fails and where a body part stays hidden, and it explains the failure under random visibility, where neighbouring frames contradict each other. The coherence boundary is therefore not an exception to the mechanism but its signature. It also predicts that unstable tracking, jittery crops, and irregular frame sampling will erode the advantage.

Temporal modelling may belong in the representation. Since GaitSet, gait research has treated temporal modelling as a problem of the gait model, addressed with set pooling, temporal modules, 3D convolutions, snippets, transformers, and motion branches (Chao et al., 2019; Fan et al., 2020; 2024a; Hou et al., 2026b;a; Huang et al., 2026a). Our results suggest a different division of labour. With the gait head fixed as an order-invariant set pooling, clip-wise encoding alone improves cross-domain recognition and makes unchanged frames more identity-discriminative. With large-scale video pretraining, temporal modelling may be better placed in the representation, leaving the gait head to aggregate. A gait head designed for clip-wise features, which ours does not attempt, is the natural next step.

Video LVMs as a foundation for gait recognition. Video LVMs are not universal replacements for image LVMs: within-domain, image LVMs still lead across model families (Table 1), and under incoherent visibility, independently encoded frames are safer. Their value lies where identity evidence is unreliable, across domains and under persistent occlusion, which a choice of foundation model by within-domain Rank-1 would miss. This regime, partial views of the body under changing conditions, is also how embodied platforms such as mobile robots observe people (Martín-Martín et al., 2023). Gait-based identity recognition there remains largely unexplored, and our results make video LVMs a natural foundation for it.

Limitations. The detailed diagnostics concern V-JEPA 2.1-B, one source dataset (CCPG), and one order-invariant gait head, and each recognition model is trained once , so we do not read margins below 1% as rankings. Occlusion is synthetic, and the assays show identity-relevant feature changes without showing that they mediate the entire recognition gain. Whether encoders with weaker temporal objectives gain less, and how context inside the encoder compares with order-aware gait heads on frame-wise features (Huang et al., 2026a; Orlova et al., 2026), remain open.

7 Related Work

Temporal modelling in gait recognition. Gait models have been built on silhouettes, skeletons, parsing maps, and 3D bodies (Chao et al., 2019; Fan et al., 2024b; Wang et al., 2024a; Zheng et al., 2022). GaitSet introduced the view of gait as a set (Chao et al., 2019), later defended by GaitAsset (Hou et al., 2025), while other models learn temporal structure with part-based temporal modules, 3D convolutions, snippets, or transformer backbones (Fan et al., 2020; Lin et al., 2022; Fan et al., 2024a; Hou et al., 2026b;a). LVM-based methods read gait from frozen image LVMs through set-style gait heads (Ye et al., 2024; 2025), and GaitMax adds a motion branch on per-frame LVM features (Huang et al., 2026a). In all of them, temporal structure is learned by the gait model; we ask what changes when it is already present in the foundation representation.

Occluded gait recognition. Occluded gait recognition registers or reconstructs occluded silhouettes, fits occlusion-aware body models, routes inputs to occlusion-specific experts, or distils knowledge from unoccluded sequences (Uddin et al., 2019; Xu et al., 2023b;a; Huang et al., 2024; Gupta & Chellappa, 2024; 2025); occluded video re-identification completes hidden regions from other frames (Hou et al., 2019; 2022), and quality-aware gait models down-weight unreliable frames (Hou et al., 2023; Wang et al., 2024b). These methods recover or discount missing evidence. We instead ask how a frozen representation organizes the evidence that remains.

What video pretraining adds. Video self-supervision learns clip-wise representations by masked reconstruction or latent prediction (Tong et al., 2022; Bardes et al., 2024; Assran et al., 2025; Wang et al., 2025a; Mur-Labadia et al., 2026). Studies of what it adds over images compare different models or separately trained image and video versions (Vélez et al., 2025; Huang et al., 2026b; Liu et al., 2026; Kodathala & Vunnam, 2025), or train new temporal modules on image-pretrained encoders (Dave et al., 2024; Orlova et al., 2026), and frame shuffling probes temporal dependence (Misra et al., 2016; Kowal et al., 2025). We instead run one frozen checkpoint through its native image and video pathways, study identity under controlled evidence degradation, and locate the effect in frozen tokens.

8 Conclusion

GaitSet-style pipelines treat gait as a set of independently encoded frames and refine how these are pooled. Running one V-JEPA 2.1 checkpoint frame-wise and clip-wise under identical training, we find that encoding frames in context improves the features being pooled: it helps most where appearance is least reliable and when a body part stays hidden, but becomes a liability when visibility changes incoherently, where independently encoded frames are safer. Video LVMs are therefore most compelling not as universal replacements for image encoders, but as a foundation for gait recognition when identity evidence is unreliable, provided that the pipeline keeps visibility coherent.

AI USE STATEMENT

We used OpenAI GPT-based tools, including ChatGPT and Codex, to assist with drafting and revising portions of the manuscript; polishing language, clarity, readability, and organization; LaTeX formatting; and retrieving or discovering potentially relevant literature and bibliographic information. All AI-generated text and suggestions were reviewed and edited by the authors, who take full responsibility for the final manuscript, citations, claims, analyses, and artifacts.

REPRODUCIBILITY STATEMENT

Section 2 describes the common gait-head architecture, datasets, evaluation protocols, and frame-wise and clip-wise encoding modes. Sections 4 and 5 specify the occlusion experiments and representation diagnostics. Appendices A–C provide computational accounting, additional robustness results, and detailed diagnostic protocols, including the fixed context-derangement seeds and identity-level bootstrap procedures.

We plan to release the code and experiment configurations.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL https://arxiv.org/abs/2404.08471.

Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):8126–8133, 2019. doi: 10.1609/aaai.v33i01.33018126.

Ishan Rajendrakumar Dave, Simon Jenni, and Mubarak Shah. No more shortcuts: Realizing the potential of temporal self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1481–1491, 2024. doi: 10.1609/aaai.v38i2.27913.

Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14213–14221, 2020. doi: 10.1109/CVPR42600.2020.01423.

Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition toward better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9707–9716, 2023. doi: 10.1109/CVPR52729.2023.00936.

Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring deep models for practical gait recognition, 2024a. URL https://arxiv.org/abs/2303.03301.

Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. Skeletongait: Gait recognition using skeleton maps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2):1662–1669, 2024b. doi: 10.1609/aaai.v38i2.27933.

Ayush Gupta and Rama Chellappa. You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5881–5890, 2024. doi: 10.1109/WACV57701.2024.00579.

Ayush Gupta and Rama Chellappa. MimicGait: A model agnostic approach for occluded gait recognition using correlational knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4757–4766, 2025.

Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7183–7192, 2019. doi: 10.1109/CVPR.2019.00735.

Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4894–4912, 2022. doi: 10.1109/TPAMI.2021.3079910.

Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang. Gait quality aware network: Toward the interpretability of silhouette-based gait recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8978–8988, 2023. doi: 10.1109/TNNLS.2022.3154723.

Saihui Hou, Chenye Wang, Aoqi Li, Jilong Wang, Liang Wang, and Yongzhen Huang. Gaitasset: In defense of regarding gait as a set. *IEEE Transactions on Information Forensics and Security*, 20: 12301–12316, 2025. doi: 10.1109/TIFS.2025.3630892.

Saihui Hou, Wenpeng Lang, Jilong Wang, Yan Huang, Liang Wang, and Yongzhen Huang. Gait transformer: End-to-end transformer backbone for gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4771–4779, 2026a. doi: 10.1609/aaai.v40i6.42479.

Saihui Hou, Chenye Wang, Wenpeng Lang, Zhengxiang Lan, and Yongzhen Huang. Gaitsnippet: Gait recognition beyond unordered sets and ordered sequences. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2508.07782.

Panjian Huang, Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Zhiqiang He, and Yongzhen Huang. Occluded gait recognition with mixture of experts: An action detection perspective. In *Computer Vision–ECCV 2024*, pp. 380–397. Springer, 2024. doi: 10.1007/978-3-031-72658-3_22.

Zhanbo Huang, Dingqiang Ye, Xiaoming Liu, and Yu Kong. Unlocking motion from large vision models with a semantic and kinematic duality for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28379–28390, 2026a.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M. Rehg. How much 3D do video foundation models encode? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. arXiv:2512.19949.

Dongyang Jin, Chao Fan, Weihua Chen, and Shiqi Yu. Exploring more from multiple gait modalities for human identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4): 4120–4128, 2025a. doi: 10.1609/aaai.v39i4.32432.

Dongyang Jin, Chao Fan, Jingzhe Ma, Jingkai Zhou, Weihua Chen, and Shiqi Yu. On denoising walking videos for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12347–12357, 2025b. doi: 10.1109/CVPR52734.2025.01152.

Sai Varun Kodathala and Rakesh Vunnam. Temporal vs. spatial: Comparing DINOv3 and V-JEPA2 feature representations for video action analysis, 2025.

Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil D. B. Bruce, Richard P. Wildes, and Konstantinos G. Derpanis. Quantifying and learning static vs. dynamic information in deep spatiotemporal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 190–205, 2025. doi: 10.1109/TPAMI.2024.3462291.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13824–13833, 2023. doi: 10.1109/CVPR52729.2023.01328.

Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *Computer Vision–ECCV 2022*, pp. 375–390. Springer, 2022. doi: 10.1007/978-3-031-20065-6_22.

Beibei Lin, Shunli Zhang, Ming Wang, Lincheng Li, and Xin Yu. Gaitgl: Learning discriminative global-local feature representations for gait recognition, 2022.

Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, Xilin Zhao, and Qingming Huang. From static to dynamic: Exploring self-supervised image-to-video representation transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. arXiv:2603.26597.

Roberto Martín-Martín, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6748–6765, 2023. doi: 10.1109/TPAMI.2021.3070543.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016*, pp. 527–544. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46448-0_32.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning, 2026. URL https://arxiv.org/abs/2603.14482.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

Svetlana Orlova, Niccolò Cavagnero, and Gijs Dubbelman. Towards data-efficient video pre-training with frozen image foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026. arXiv:2605.19137.

Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1054–1063, 2023. doi: 10.1109/CVPR52729.2023.00108.

Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q. Huang, and Liang Wang. A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(2):270–292, 2025. doi: 10.1109/TBIOM.2024.3486345.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732.

Md. Zasim Uddin, Daigo Muramatsu, Noriko Takemura, Md. Atiqur Rahman Ahad, and Yasushi Yagi. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Transactions on Computer Vision and Applications*, 11:9, 2019. doi: 10.1186/s41074-019-0061-3.

Pedro Vélez, Luisa F. Polanía, Yi Yang, Chuhan Zhang, Rishabh Kabra, Anurag Arnab, and Mehdi S. M. Sajjadi. From image to video: An empirical study of diffusion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16948–16958, 2025.

Chenting Wang, Yuhan Zhu, Yicheng Xu, Jiange Yang, Ziang Yan, Yali Wang, Yi Wang, and Limin Wang. Internvideo-next: Towards general video foundation models without video-text supervision, 2025a.

Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, and Yongzhen Huang. Gaitparsing: Human semantic parsing for gait recognition. *IEEE Transactions on Multimedia*, 26:4736–4748, 2024a. doi: 10.1109/TMM.2023.3325962.

Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, Yongzhen Huang, Peipei Li, and Shibiao Xu. QAGait: Revisit gait recognition from a quality perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5785–5793, 2024b. doi: 10.1609/aaai.v38i6.28391.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13259–13269, 2025b.

Chi Xu, Yasushi Makihara, Xiang Li, and Yasushi Yagi. Occlusion-aware human mesh model-based gait recognition. *IEEE Transactions on Information Forensics and Security*, 18:1309–1321, 2023a. doi: 10.1109/TIFS.2023.3236181.

Chi Xu, Shogo Tsuji, Yasushi Makihara, Xiang Li, and Yasushi Yagi. Occluded gait recognition via silhouette registration guided by automated occlusion degree estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3191–3201, 2023b. doi: 10.1109/ICCVW60793.2023.00344.

Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 200–210, 2024. doi: 10.1109/CVPR52733.2024.00027.

Dingqing 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, 2025.

Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition*, pp. 441–444, 2006. doi: 10.1109/ICPR.2006.67.

Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20196–20205, 2022. doi: 10.1109/CVPR52688.2022.01959.

Jinkai Zheng, Xinchen Liu, Boyue Zhang, Chenggang Yan, Jiyong Zhang, Wu Liu, and Yongdong Zhang. It takes two: Accurate gait recognition in the wild via cross-granularity alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8786–8794, 2024. doi: 10.1145/3664647.3680577.

Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. Cross-covariate gait recognition: A benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7): 7855–7863, 2024. doi: 10.1609/aaai.v38i7.28621.

APPENDIX: ADDITIONAL RESULTS AND ANALYSIS

This appendix follows the evidence chain of the main paper. Section A reports computational cost. Section B tests whether the robustness trend extends across gait conditions and depends on how visible evidence is organized over time. Section C examines whether the corresponding effects are already present in frozen tokens and whether they carry identity information.

A COMPUTATIONAL PROFILE

We report model size and forward cost to test whether the recognition trends can be reduced to capacity or computation alone.

Table 4: Total system parameters and theoretical cost of a 32-frame inference pass. Frame-wise costs sum 32 independent forwards. GFLOPs characterize forward computation rather than training speed.

Backbone	Exec.	Resolution	Params (M)	GFLOPs
DINOv2-S	Frame	448×224	142.2	1460.0
DINOv2-L	Frame	448×224	430.0	6505.0
InternVideo-Next-B	Clip	448×224	142.0	6719.6
V-JEPA-B	Clip	512×256	128.4	2128.3
V-JEPA-L	Clip	512×256	346.8	5976.4

The profile uses batch size one and counts both matrix products in fused scaled-dot-product attention. The cost ranking is not monotonic with recognition performance: InternVideo-Next-B is the most expensive model but does not lead the comparison, while V-JEPA-L uses fewer parameters and GFLOPs than DINOv2-L. Model scale and forward cost therefore do not explain the source–target ordering. Actual runtime also depends on activation storage, backward computation, kernels, and data loading.

B ADDITIONAL ROBUSTNESS RESULTS

B.1 FULL PERSISTENT-OCCLUSION RESULTS

Figure 4 expands the primary CL comparison to UP, DN, BG, and Avg., while keeping body location and severity separate. It therefore tests whether the main-paper trend depends on a particular mask or gait condition. All values use gait Rank-1 excluding identical-view cases.

The loss varies markedly with body location and gait condition, showing that body regions contribute different amounts of identity evidence. Despite this heterogeneity, the clip-wise executions generally separate more favorably from their frame-wise references as the persistent region grows. The robustness gain is therefore not a constant clean-score offset; it emerges under sustained, severe evidence loss and is not driven by one mask location or by CL alone.

B.2 TIME-VARYING OCCLUSION

Table 5 reports two 40% time-varying masks that remove the same amount of evidence per frame but organize the surviving evidence differently. Smooth moves the band coherently, whereas Random resamples its vertical position independently in each frame. All verified rows use the same ImageNet-mean RGB fill and a clean gallery.

Smooth visibility preserves the clip-wise advantage over matched frame-wise V-JEPA, whereas Random visibility reverses it. DINOv2-S improves from Smooth to Random, while frame-wise V-JEPA is unchanged, so Random is not uniformly harder. Instead, the reversal is specific to clip-wise execution when visibility changes incoherently. Evidence reappearance is therefore insufficient by itself; the temporal organization of the surviving evidence also matters. Because all rows use one fill color, these results do not establish invariance to mask appearance.

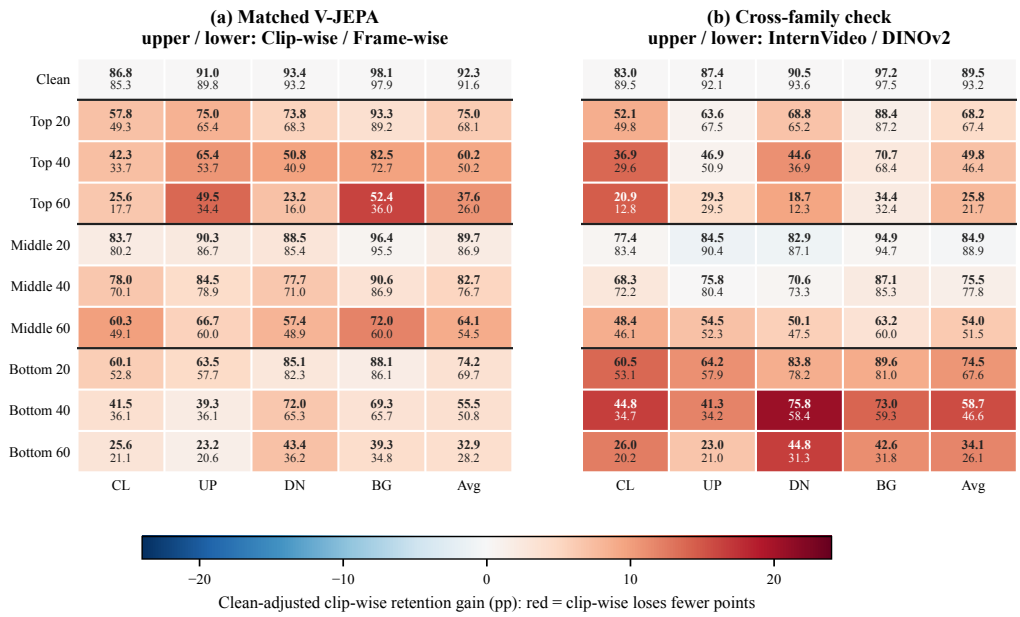

Figure 4: Complete CCPG results under persistent probe-only occlusion. The masked band covers 20%, 40%, or 60% of body height at the top, middle, or bottom and remains fixed throughout the sequence; all gallery sequences remain clean.

Table 5: CCPG gait Rank-1 (%) under time-varying 40% probe-only occlusion. Smooth moves the band coherently, whereas Random resamples its position independently in each frame. Avg. averages CL, UP, DN, and BG. All rows use RGB fill $(124, 116, 104)$.

Backbone	Execution	Mask trajectory	CL	UP	DN	BG	Avg.
DINOv2-S	Frame	Smooth	80.6	86.2	88.3	94.2	87.3
DINOv2-S	Frame	Random	82.4	87.7	89.6	94.9	88.7
V-JEPA-B	Frame	Smooth	77.8	84.4	89.2	96.0	86.8
V-JEPA-B	Frame	Random	77.8	84.4	89.2	96.0	86.8
V-JEPA-B	Clip	Smooth	79.4	85.8	89.3	95.9	87.6
V-JEPA-B	Clip	Random	70.1	78.6	85.4	93.1	81.8

C Additional Representation Diagnostics

C.1 Raw-Token Measurement

At each native frozen-token coordinate, we use the cosine drift defined in Eq. (2) of the main paper and compute relative-ℓ_2 as $\|\tilde{\mathbf{z}} - \mathbf{z}\|_2 / \|\mathbf{z}\|_2$. Both are measured before pooling, layer fusion, the foreground adapter, CNN processing, or the gait readout, thereby isolating the frozen representation's response to persistent corruption.

The opposite directions at visible and overwritten locations are the key phenomenon. Clip-wise tokens drift less where evidence survives, yet respond more strongly where the input is overwritten. Joint encoding therefore does not make the representation globally inert or uniformly smooth; it selectively stabilizes surviving evidence while remaining sensitive to actual input changes. This pattern is consistent with the recognition robustness, but it does not imply reconstruction of hidden pixels.

C.2 Fixed-Observation Protocol and Identity Assay

We fix anchor tubelet 8 (zero-based frames 16–17), including its pixels and absolute position, and derange only the other 15 tubelets with seeds 20260711, 20260712, and 20260713. The anchor is bitwise unchanged. As defined in Eq. (3) of the main paper, Δm compares its identity margin under natural context with the mean under the three derangements. Because the fixed readout receives

Table 6: Frozen-token drift for the matched V-JEPA-B executions, averaged over four preselected layers and all nine persistent masks. All measurements precede any trained gait-specific operation.

Input region	Metric	Clip	Frame	Clip−Frame
Still visible	Cosine	0.014105	0.016876	-0.002771
Overwritten	Cosine	0.345607	0.274288	$+0.071319$
Still visible	Rel.-ℓ_2	0.125636	0.141136	-0.015500
Overwritten	Rel.-ℓ_2	0.806533	0.718800	$+0.087733$

Table 7: Fixed-observation identity assay on clean probes ($\Delta m \times 10^3$). Deranged values average three fixed seeds; intervals use 10,000 identity-level bootstrap resamples. Positive values indicate that natural context increases the identity margin of the unchanged anchor.

Protocol	Frame Δm	Clip Δm	Clip 95% CI
CL	0.000	2.762	[2.324, 3.201]
UP	0.000	2.664	[1.952, 3.441]
DN	0.000	1.634	[0.903, 2.368]
BG	0.000	3.258	[2.596, 3.935]
Avg.	0.000	2.579	[2.248, 2.911]

only the anchor representation, the assay tests whether the surrounding clip has already changed its identity value inside the frozen encoder.

Frame-wise V-JEPA is exactly invariant because it encodes the anchor independently. Clip-wise Δm is positive for every CCPG protocol, so the effect is not confined to CL. Natural context consistently increases the separation of the unchanged anchor from competing identities. Since the readout never sees the surrounding tubelets, this identity-relevant change is already present in the frozen clip-conditioned token.

C.3 Occlusion Extension of the Identity Assay

We extend the fixed-observation assay to persistent occlusion using the same anchor tubelet and the same three context derangements as in Section C.2. Within each natural–deranged comparison, the anchor pixels, absolute position, occlusion mask, clean gallery, and gait readout remain fixed; only the ordering of the other 15 tubelets changes. For body location b and severity r, we report

$$\Delta\Delta m_{b,r} = \Delta m_{b,r} - \Delta m_{\text{clean}}, \tag{4}$$

where Δm measures the identity-margin benefit of natural rather than deranged temporal context. Positive $\Delta\Delta m$ indicates that natural context provides a larger identity benefit under the specified occlusion than on clean probes.

The effect is strongly location dependent. Top occlusion increases $\Delta\Delta m$ at all three severities, while middle occlusion clearly exceeds the clean-context benefit only at 60%. In contrast, the confidence intervals under bottom occlusion include zero, and the 60% estimate becomes negative. The contextual benefit is therefore not a generic consequence of corrupting the input; it depends on which body evidence remains available to the jointly encoded clip.

One plausible explanation is that 40% and 60% bottom masks remove much of the periodic lower-body motion and may also intercept portions of the arm-swing trajectory. These masks therefore leave less coherent motion-rich evidence for clip-wise contextualization. Top masks, by contrast, preserve the leg dynamics, while middle masks can retain some distal lower-body motion. Removing these cues may explain why the additional identity benefit relative to clean probes disappears, or becomes negative, under severe bottom occlusion.

This interpretation remains tentative because the current intervention does not isolate the contributions of individual body parts or directly measure the amount of motion information retained by each mask. Moreover, negative $\Delta\Delta m$ does not by itself imply a reversal in overall recognition accuracy; it means only that the natural-context benefit for the fixed anchor is smaller than on clean probes. The experiment therefore supports location-dependent contextual compensation, rather than establishing that body-part motion mediates the full Rank-1 robustness advantage.

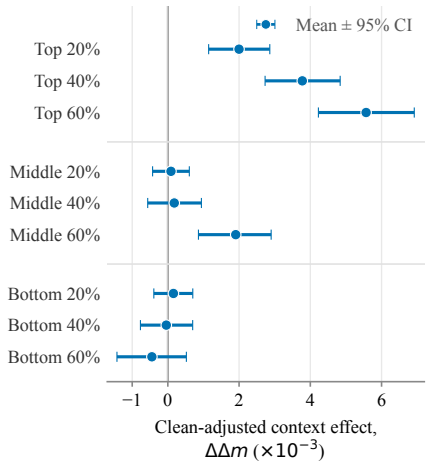

Figure 5: Occlusion extension of the fixed-observation identity assay. Points show $\Delta\Delta m$, the change in natural-context benefit relative to clean probes; bars are identity-bootstrap 95% confidence intervals. Body locations are reported separately.

D Alternative Explanations and Controls

Table 8 lists simpler explanations of the clip-wise advantage and the controls that address them. None accounts for the full pattern across §3–§5; what remains is consistent with the temporal prior of video pretraining discussed in §6. This account predicts that unstable tracking, jittery crops, or irregular frame sampling will erode the advantage, and that encoders with weaker temporal objectives will show a smaller one.

Table 8: Alternative explanations and the controls that address them (matched V-JEPA 2.1-B unless noted).

If the advantage came from …	Control	Observation
different weights, capacity, or compute	same checkpoint; cost accounting	gap persists within one checkpoint; V-JEPA-L is cheaper than DINOv2-L
the trained gait head	token drift before the gait head	encoding mode changes frozen tokens (§5.2)
re-exposure of hidden regions	persistent vs. Smooth vs. Random bands	largest gain without re-exposure; reversed under incoherent re-exposure (§4)
filling in occluded content	drift at overwritten positions	clip-wise tokens change 26% *more* there
the image vs. video tokenizer	fixed-anchor assay; frame shuffles	anchor tokenization is fixed, yet $\Delta m > 0$; -38.6 CL points under full shuffle