frontiers in Imaging

TYPE Original Research
PUBLISHED 03 June 2026
DOI 10.3389/fimag.2026.1846329

OPEN ACCESS

EDITED BY
Alessandro Piva,
University of Florence, Italy

REVIEWED BY
Yunxue Shao,
Nanjing Tech University, China
Ruili Shi,
Southeast University, China

*CORRESPONDENCE
Pascal Benschop
P.Benschop@tudelft.nl

RECEIVED 02 April 2026
REVISED 01 May 2026
ACCEPTED 11 May 2026
PUBLISHED 03 June 2026

CITATION
Benschop P, van Gemert J, Mense JP and Dauwels J (2026) Motion representations for privacy-aware cross-domain action recognition. *Front. Imaging* 5:1846329. doi: 10.3389/fimag.2026.1846329

Motion representations for privacy-aware cross-domain action recognition

Pascal Benschop[1,2]*, Jan van Gemert[1], Jelte P. Mense[2] and Justin Dauwels[1]

[1]Department of Computer Science, Delft University of Technology, Delft, Netherlands, [2]National Policelab AI & Model-Driven Decisions Lab, Delft, Netherlands

Video captured for action recognition often contains sensitive appearance cues such as faces, skin color, and clothing. Models trained on such data may exploit these cues rather than the underlying motion, raising privacy concerns in real-world deployment. In this work, we study action recognition under a motion-focused constraint: the model receives only motion representations that capture pixel displacement over time, while reducing appearance cues that expose identity or scene context. We focus on motion-history images and optical flow as learning-free representations that reduce identifiable appearance information while retaining action recognition accuracy. Our motion I3D model achieves approximately 31% and 52% zero-shot top-1 accuracy on HMDB-51 and UCF-101, respectively, outperforming non-CLIP direct-transfer baselines trained on Kinetics-400 despite operating without any appearance input. In 16-shot adaptation, the same model reaches 52% and 83% top-1 accuracy. In the domain adaptation setting on TP-HMDB↔TP-UCF, our motion-focused models achieve higher action recognition accuracy than prior privacy-preserving methods. Sensitive attribute predictability is reduced relative to RGB by a comparable margin, without requiring a learned privacy filter. On PA-HMDB51, optical flow is the strongest motion representation for privacy preservation, approaching chance level for skin-color prediction and remaining below RGB on most privacy attributes, indicating that motion representations retain useful action information while exposing less personal information.

KEYWORDS
action recognition, cross-domain, domain adaptation, optical flow, privacy

1 Introduction

Action recognition is increasingly deployed in sensitive real-world environments—hospitals, care homes, sports facilities, and public spaces—where video data may be captured without explicit consent or where retention of identifiable visual information is legally or ethically undesirable. Standard RGB video exposes faces, skin tone, clothing, and scene context, all of which can be used to identify individuals even when action labels are the only intended output. This motivates a practical question: can action recognition remain useful when appearance is removed from the input?

We study this question through the lens of learning-free motion representations, specifically optical flow and Motion History Images (MHI) (Bobick and Davis, 2001). Unlike privacy filters that attempt to suppress sensitive information via learning-based transformations, these representations are computed using deterministic algorithms that

inherently suppress much of the texture, color, and fine detail present in RGB frames. Optical flow provides the strongest suppression, while MHI can still retain coarse shape and silhouette information. Both are simple and practical alternatives for privacy-aware deployment, and their combination with modern CLIP-aligned training has not been systematically evaluated in cross-domain and privacy-sensitive settings.

Our goal is not to argue that appearance is never useful, nor to claim that motion-only inputs should replace RGB in all settings. Instead, we aim to better understand the trade-off that emerges when appearance is reduced. In particular, we ask how much action-discriminative information is preserved in motion-focused inputs, how well such representations transfer across domains, and to what extent they reduce sensitivity to appearance-related attributes. This framing is especially relevant in cross-domain and privacy-sensitive settings, where methods that depend strongly on appearance may be less reliable or less desirable.

Based on this motivation, we study the following research questions:

- **RQ1:** How much cross-domain action recognition accuracy is retained when models are trained on motion-focused representations instead of RGB video?
- **RQ2:** To what extent do motion representations reduce sensitivity to appearance-related attributes such as skin color, compared to RGB-based inputs?

We hypothesize that motion representations yield lower absolute accuracy than RGB-based methods, particularly for actions that depend strongly on objects or scene context. Among the motion representations considered, optical flow is expected to provide the strongest appearance suppression, while combining it with MHI is expected to recover some recognition accuracy at a modest privacy cost. Overall, we expect motion-focused representations to be less sensitive to appearance-related attributes than RGB, while still retaining cross-domain action recognition accuracy, albeit below RGB-based methods.

The primary contribution of this paper is a practical system design and empirical study. We combine learning-free motion representations with a CLIP-aligned training pipeline and evaluate the resulting privacy–utility trade-off systematically across zero-shot, few-shot, and domain-adaptation settings. Our design choices stem from existing approaches, the combination of these results in an efficient model that is able to keep up with RGB models in similar circumstances, while significantly reducing privacy leakage.

2 Related work

2.1 Appearance-heavy action recognition

Modern action recognition is dominated by RGB-based models that learn both motion and appearance cues. 3D convolutional networks and factorized variants such as (2+1)D CNNs learn spatiotemporal features directly from RGB video (Tran et al., 2018; Carreira and Zisserman, 2017), while transformer-based models further expand temporal modeling capacity (Arnab et al., 2021; Bertasius et al., 2021). Vision–language pretraining has made pretrained appearance representations especially attractive for transfer, and recent methods such as ViFi-CLIP and TC-CLIP adapt CLIP-like encoders to video by injecting temporal structure while retaining pretrained appearance semantics (Rasheed et al., 2023; Kim et al., 2024). Although effective, these pipelines are highly dependent on appearance, scene context, and identity-related cues. Recent work confirms that video models can rely heavily on static shortcuts such as background, clothing, and scene context, which degrades cross-domain accuracy when those cues shift between training and deployment (Li et al., 2023a; Zhai et al., 2023). We contrast with these methods by removing appearance at the input level rather than attempting to suppress shortcuts after the fact.

2.2 Motion-focused action recognition

Suppressing appearance in favor of motion has a history in action recognition. Dense trajectory methods and motion boundary histograms achieved strong action recognition accuracy through explicit motion modeling (Wang and Schmid, 2013). Two-stream architectures demonstrated that optical flow is highly informative as a complementary input to RGB (Simonyan and Zisserman, 2014). More recently, Appearance-Free Action Recognition enforced models to focus on pure motion by replacing RGB in a two-stream network with warped noise, demonstrating competitive accuracy without appearance cues (Ilic et al., 2022). TranSVAE disentangles appearance and motion factors in latent space, showing that motion representations generalize more favorably under domain shift (Wei et al., 2023). Efficient recognition from compressed video streams uses coarse motion vectors rather than full RGB frames (Wu et al., 2018; Shou et al., 2019). We build on this line of work by studying how well features from motion representations transfer across domains.

2.3 Privacy in action recognition

Pose and skeleton sequences offer a privacy-friendlier alternative to RGB by abstracting away appearance (Yan et al., 2018; Chi et al., 2022; Ren et al., 2024), but depend on reliable pose estimation, which degrades for small, occluded, or crowded subjects (Wang et al., 2021; Fan and Chowdhury, 2025; Wei et al., 2024). Privacy-preserving video encoders take a different approach, learning to transform raw frames into anonymized representations that retain action-relevant information while suppressing privacy-sensitive attributes (Kumawat and Nagahara, 2022; Li et al., 2023b; Wu et al., 2022; Xia et al., 2025). The need for such methods is underscored by fairness analyses of surveillance datasets, which reveal demographic imbalances in skin color and gender across action classes, confirming that appearance cues in video carry sensitive attribute information (Pastaltzidis et al., 2022). Despite bypassing RGB entirely, we explicitly evaluate how much sensitive attribute information remains accessible in our motion representations, acknowledging that even appearance-reduced inputs may retain residual privacy-relevant cues.

2.4 Privacy-preserving sensors

An orthogonal line of work replaces RGB with sensors that inherently capture less appearance information. Depth-based action recognition avoids reliance on color and texture by using range data and has been applied to home activity monitoring and fall detection (Wang et al., 2018; Sánchez-Caballero et al., 2022). Thermal and infrared imaging preserve temporal activity under low illumination while suppressing most fine appearance detail (Batchuluun et al., 2019). Event cameras output asynchronous brightness-change events rather than frames, and have been explored for low-latency, privacy-friendly gesture and action recognition (Amir et al., 2017; Innocenti et al., 2021; Bi et al., 2020). These sensors provide hardware-level privacy but require dedicated hardware and large per-sensor labeled datasets. Our approach is complementary: we keep commodity RGB cameras and apply deterministic motion extraction at the software level, while the training framework itself (Section 3) is sensor-agnostic and can be applied to depth, infrared, or accumulated event representations with minimal modification, as discussed in Section 5.

3 Methodology

Our input representation is deliberately learning-free and computationally lightweight. Learning-free extraction is motivated by the cross-domain setting: the target domain does not provide privacy-attribute labels, so any learned anonymization filter would need to be retrained or validated per deployment, with no guarantee that privacy suppression transfers. Computational efficiency is motivated by practical deployment constraints: action recognition on surveillance infrastructure often involves many simultaneous camera feeds, and processing must remain feasible on modest hardware without dedicated preprocessing pipelines. Together, these constraints lead to deterministic motion representations and an asymmetric allocation of the spatiotemporal budget between their two components.

We consider common video representations for action analysis, including RGB, optical flow, motion-history templates, and pose/skeletons. RGB is a strong action-recognition baseline, but it preserves rich appearance cues and therefore exposes more identity-related information than motion-focused alternatives (Carreira and Zisserman, 2017; Dave et al., 2022). Optical flow suppresses much of the static appearance while retaining motion patterns useful for recognition, and part of its effectiveness has been attributed to its relative invariance to appearance (Simonyan and Zisserman, 2014; Sevilla-Lara et al., 2017). Motion History Images (MHI) provide a compact summary of motion over time and are widely used as a simple and robust temporal representation (Bobick and Davis, 2001). In contrast, pose/skeleton pipelines depend on an upstream detector or pose estimator, which can become unreliable in low-quality footage with blur, low resolution, occlusion, poor lighting, or small subjects (Osokin, 2018). An analysis of pose estimation quality on the UCF-Crime dataset (Sultani et al., 2018) is provided in Supplementary Section 1. Pose detections were unusable across almost all videos.

Guided by these observations, our input representation is deliberately motion-focused and learning-free. Given grayscale frames $\{G_t\}_{t=1}^{T}$, where T is the number of frames in a video and $\{G_t\}$ denotes the grayscale frame at time index t, we construct two fixed-size motion inputs. First, we sample S frame pairs approximately uniformly across the full video duration and compute dense Farnebäck optical flow (Farnebäck, 2003) between consecutive grayscale frames. The resulting flow sequence is

$$\mathbf{F} = \{\phi(G_{i_s-1}, G_{i_s})\}_{s=1}^{S}, \tag{1}$$

where ϕ denotes dense optical flow and i_s is the selected frame index for sample s. The two motion channels are intentionally stored at different spatiotemporal resolutions: flow captures short-horizon displacement and benefits more from temporal density than spatial detail, whereas each MHI snapshot already aggregates motion over a horizon of multiple frames and requires higher spatial resolution to preserve shape. Inspired by YOLO-I3D (Luo et al., 2024), we use $S = 128$ flow fields of spatial size 112×112, with two channels corresponding to horizontal and vertical motion. Lower resolution has another benefit: it reduces computational and I/O cost. We use dense Farnebäck optical flow because it is learning-free, computationally efficient relative to modern learned flow estimators, and can be applied to a range of input modalities, provided that consecutive frames exhibit sufficiently consistent local structure for motion estimation.

Second, we maintain a motion history image $\mathbf{H}_t$ driven by the frame-difference mask

$$D_t(x) = \mathbb{I}\big(|G_t(x) - G_{t-1}(x)| > \delta\big), \tag{2}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, which is 1 if the condition is true and 0 otherwise. This MHI is updated with the following rule

$$H_t(x) = \begin{cases} \tau, & D_t(x) = 1, \\ \max(0, H_{t-1}(x) - 1), & D_t(x) = 0, \end{cases} \tag{3}$$

We set the temporal horizon to $\tau = 25$ and threshold $\delta = 15$, which performed best in our ablation (see Section 4.5). This can be interpreted as a temporally decaying memory of recent motion, where more recent motion produces higher-intensity values. Since each MHI snapshot summarizes motion over a longer temporal window than a single optical flow field, fewer snapshots are needed to cover the same clip duration. We therefore store 32 snapshots at resolution 224×224, which we denote collectively by $\mathbf{H}$. An example of the resulting flow representation, together with motion history images, is shown in Figure 1.

For experiments that use both optical flow and MHI as inputs, the two motion channels feed a two-stream I3D model with late fusion; see Figure 2 for an overview of this configuration. The upper-left branch visualizes MHI inputs, while the lower-left branch shows optical-flow inputs at higher temporal resolution. In the optical-flow-only setting, the same training pipeline is used with only the flow stream enabled.

Let $f_{\text{flow}}(\cdot)$ and $f_{\text{mhi}}(\cdot)$ denote the corresponding I3D encoders, and let $g(\cdot)$ denote a linear projection head that maps the pooled encoder output to the shared embedding space. In the optical-flow-only setting, the fused representation is

$$\mathbf{v} = g\big(f_{\text{flow}}(\mathbf{F})\big). \tag{4}$$

FIGURE 1
Top: optical flow estimated with Farnebäck and overlaid on RGB frames, where green arrows indicate the direction and relative magnitude of motion between adjacent frames. **Bottom**: corresponding motion history images (MHIs), which suppress appearance while retaining a coarse motion silhouette. In contrast, the Farnebäck flow field is noisy and low fidelity, making the actor's silhouette much less discernible. Adapted from Kinetics-400, with overlayed arrows to show the optical flow, and a conversion to motion history images in the second row. Source Kinetics: https://research.google/pubs/the-kinetics-human-action-video-dataset/.

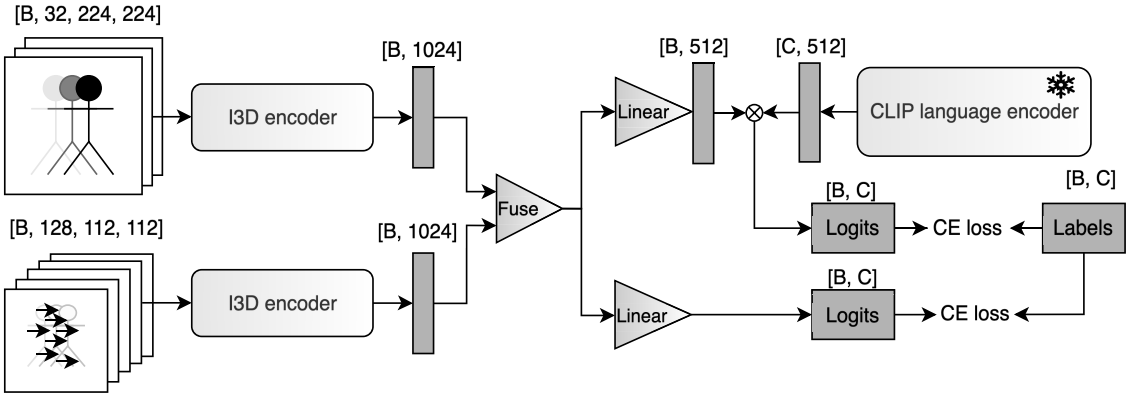

FIGURE 2
Two-stream I3D model for contrastive training with MHI and optical flow. After fusion, the video feature is linearly projected and matched to frozen CLIP text embeddings via the ⊗ (dot product) operator, producing class logits used in the cross-entropy loss. The class head (located in the lower right branch) is also optimized directly using cross-entropy loss to estimate the kinetics class for each video.

In the combined setting, the two stream embeddings are averaged before projection,

$$\mathbf{v} = g\left(\frac{f_{\text{flow}}(\mathbf{F}) + f_{\text{mhi}}(\mathbf{H})}{2}\right). \tag{5}$$

The final video embedding is then ℓ_2-normalized as

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}. \tag{6}$$

ℓ_2 normalization projects all embeddings onto the unit hypersphere, ensuring that similarity between embeddings is measured purely by angle rather than magnitude.

The main experiments use I3D as the encoder backbone. A lightweight alternative using X3D-XS (Feichtenhofer, 2020) in the two-stream approach, inspired by Ilic et al. (2022), is evaluated in Supplementary Section 5. The X3D-E2S OF+MHI model offers a 42× reduction in GFLOPs and practical CPU-speed inference at a modest accuracy cost.

Motion embeddings are aligned to a CLIP-derived text prototype space to enable transfer without training a dataset-specific classifier. Following TC-CLIP (Kim et al., 2024) and related approaches, we augment each class label with 5 LLM-generated short descriptions, resulting in a total of 400 × (5+1) = 2,400 text entries. These descriptions provide richer semantic context, which is beneficial given the fine-grained and specific nature of many Kinetics class labels. All text entries are encoded using a pretrained CLIP text encoder, which is kept frozen throughout training.

CLIP text features are trained against RGB appearance and are therefore not optimal targets for motion-only video representations. At the same time, the CLIP text geometry encodes the semantic co-location of related labels, which is precisely what enables zero-shot transfer. We therefore adapt the text features rather than replace them: a lightweight residual adapter, implemented as a small MLP with a residual connection initialized near identity, shifts the text features toward the motion domain while preserving CLIP's label-space structure. To keep this drift

controlled, we regularize the adapted text embeddings toward the frozen embeddings with a squared-distance penalty.

Supervision is provided through a multi-positive video-to-text objective: for a video of class y_i, the target is a soft distribution $q^{(i)}$ over the full text bank. We optimize the resulting soft-target cross-entropy over all text entries:

$$\mathcal{L}_{\text{clip\_CE}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{N} q_j^{(i)} \log \frac{\exp\big(s\,\langle \hat{\mathbf{v}}_i, \mathbf{t}_j\rangle\big)}{\sum_{k=1}^{N}\exp\big(s\,\langle \hat{\mathbf{v}}_i, \mathbf{t}_k\rangle\big)}, \tag{7}$$

where B is the batch size, N is the total number of text entries in the bank, $\hat{\mathbf{v}}_i$ is the normalized video embedding of sample i, $\mathbf{t}_j$ is the j-th normalized text embedding, $q^{(i)}$ is the soft target distribution for sample i, and s is the inverse-temperature scale. Class labels are assigned a weight of 50%, with the remaining 50% distributed equally across the five descriptions.

We additionally train an auxiliary classification head, a linear layer applied directly to the raw fused embedding $\mathbf{v}$, with standard cross-entropy over the 400 Kinetics class labels. This head is used in fully supervised experiments and provides a direct discriminative gradient that stabilizes encoder training, particularly early in optimization when the alignment objective over 2400 fine-grained text descriptions is relatively weak. The two objectives are complementary: the CE head encourages source-domain class discrimination, while $\mathcal{L}_{\text{clip\_CE}}$ aligns a separate projection of the video features with the language-defined semantic space.

To regularize the embedding space geometry, we apply representation mixing during training. Specifically, mixing is performed in the embedding space after the encoder, where we interpolate both the video embedding and the corresponding text prototype. A cosine consistency loss encourages the mixed video embedding to match the mixed text target, imposing a linearity prior on the embedding space that benefits zero-shot transfer. The final loss is given by:

$$\mathcal{L} = \lambda_{\text{clip\_CE}}\,\mathcal{L}_{\text{clip\_CE}} + \lambda_{\text{CE}}\,\mathcal{L}_{\text{CE}} + \lambda_{\text{mix}}\,\mathcal{L}_{\text{mix}} + \lambda_{\text{reg}}\,\mathcal{L}_{\text{text\_reg}}. \tag{8}$$

At inference, zero-shot recognition is performed by comparing a test clip embedding $\hat{\mathbf{v}}$ against text prototypes $\{\hat{\mathbf{t}}_c\}$ and selecting the maximum cosine similarity:

$$\hat{y} = \arg\max_c \langle \hat{\mathbf{v}}, \hat{\mathbf{t}}_c\rangle. \tag{9}$$

4 Results

All models are implemented in PyTorch and trained on NVIDIA L40 GPUs. Motion representations are precomputed and stored as compressed binary files prior to training. This process, including all experimental details, can be found in the Supplementary material.

The experiments are organized into two parts. The first part focuses on cross-dataset action recognition across three transfer regimes: zero-shot, few-shot, and domain adaptation. These settings allow us to assess how much action-relevant information is retained when the model operates on motion-focused inputs alone. The second part evaluates privacy more directly, by measuring how much sensitive appearance-related information can still be inferred from the learned representations. Together, these experiments provide a structured view of the privacy–utility trade-off across multiple transfer settings.

4.1 Zero-shot transfer

Zero-shot transfer represents the most restrictive setting, as no target-domain labels are available during training. Models are trained on Kinetics-400 (Kay et al., 2017), a large-scale action recognition dataset containing approximately 240k training videos spanning 400 human action classes, and evaluated directly on UCF-101 (Soomro et al., 2012) and HMDB-51 (Kuehne et al., 2011). UCF-101 covers 101 action classes across roughly 13k videos, while HMDB-51 contains 51 action classes across approximately 7k videos. Both are widely used benchmarks covering a broad variety of human actions, and are the standard evaluation targets for the zero-shot baselines we compare against. Following prior CLIP-based approaches, both class labels and LLM-generated class descriptions are used during training. This experiment tests whether motion-focused representations, aligned with text, capture sufficient semantic structure to generalize to unseen datasets without adaptation. We evaluate optical flow and OF+MHI as the two primary motion operating points; MHI-only is excluded because we assume it retains sufficient silhouette information, thereby weakening privacy guarantees without providing a compensating utility benefit.

Table 1 shows that our motion-only model achieves lower zero-shot accuracy than CLIP-based methods on both HMDB-51 and UCF-101. This gap is expected. In contrast to baseline zero-shot action recognition models, such as ViFi-CLIP, MAXI, and TC-CLIP, our approach does not benefit from a large pretrained RGB-language model. The non-CLIP direct-transfer baselines and our motion models use only the Kinetics-400 dataset. An apples-to-apples comparison is therefore only possible against methods in the lower half of the table, which share our pretraining regime and operate at a comparable scale. Both configurations of our motion model achieve similar accuracy, suggesting that language alignment quality, rather than the richness of motion inputs, is the primary bottleneck for zero-shot transfer. At the same time, our model outperforms earlier non-CLIP zero-shot baselines, such as R(2+1)D-18 trained on Kinetics-400 from Brattoli et al. (2020).

Our method lags behind stronger non-CLIP transfer methods, such as SVT-96, on UCF-101. We attribute this gap in part to differences in architecture and pretraining scale, as SVT-96 leverages both a transformer backbone and the larger Kinetics-600 dataset, which contains substantially more videos and classes than Kinetics-400. Training on Kinetics-600 would significantly increase computational and storage requirements, which becomes particularly restrictive in our setting, where additional motion representations must also be precomputed and stored. Despite the overall gap to CLIP-based methods, the improvement over non-CLIP baselines suggests that motion-only inputs retain useful action-discriminative information, even under comparatively weak text alignment.

TABLE 1 Zero-shot Top-1 accuracy (%).

Method	Transfer source	HMDB-51	UCF-101	K600
CLIP middle frame	CLIP	41.5 ± 0.5	64.1 ± 0.8	–
ViFi-CLIP (Rasheed et al., 2023)	CLIP + K400	51.3	76.8	71.2
MAXI (Lin et al., 2023)	CLIP + K400	52.3 ± 0.7	78.2 ± 0.8	71.5 ± 0.8
TC-CLIP (LLM) (Kim et al., 2024)	CLIP + K400	56.0 ± 0.3	85.4 ± 0.8	78.1 ± 1.0
Kinetics direct transfer (non-CLIP)				
R(2+1)D-18 (Brattoli et al., 2020)	K400	22.5	44.5	–
R(2+1)D-18	K700	25.6	49.7	–
SVT-96 (Doshi and Ylmaz, 2023)	K600	40.2	68.3	–
I3D OF (ours)	K400	30.65 ± 0.9	52.2 ± 1.4	–
I3D OF + MHI (ours)	K400	30.63 ± 0.1	52.6 ± 0.3	–

The transfer source indicates the pretraining done before direct evaluation on downstream datasets. CLIP+K400 denotes CLIP-based models further adapted on Kinetics-400. Both OF and MHI-OF perform better than an older baseline trained on the Kinetics-400 dataset, yet are limited by our language alignment pretraining.

4.2 Few-shot transfer

Few-shot transfer relaxes the zero-shot setting by allowing a small number of labeled target examples. We fine-tune on $K \in \{8, 16\}$ labeled examples per class on HMDB-51, UCF-101, and Something-Something v2 (SSv2) (Goyal et al., 2017). SSv2 contains 174 fine-grained hand-object interaction classes and is specifically designed to require temporal reasoning rather than appearance recognition, making it a challenging complement to UCF-101 and HMDB-51. We use the same manifests as TC-CLIP (Kim et al., 2024) for reproducibility.

During fine-tuning, we freeze the entire I3D backbone, including batch normalization statistics, and train only the projection and classification heads. This is intentional: with only $K \times C$ total training samples (e.g., $8 \times 51 = 408$ samples for HMDB-51 at $K = 8$), updating the backbone risks catastrophic forgetting and overfitting. The frozen backbone thus acts as a fixed feature extractor, and performance directly reflects the quality of pretrained representations.

Training runs for 50 epochs with AdamW (lr=2×10^{-4}, weight decay=10^{-3}, cosine schedule), a batch size of 16, and 100 warm-up steps. We use data augmentation, mixup, and representation mixing (Section 3) to regularize the head without destabilizing the frozen encoder. We also fine-tune a Kinetics-pretrained R(2+1)D (Tran et al., 2018) model on the same splits as an RGB baseline, allowing us to partially isolate the effect of the motion-only constraint from differences in architecture and pretraining strategy.

Table 2 shows that our model adapts consistently across all three benchmarks, with performance improving from $K = 8$ to $K = 16$ in every case. The strongest results are obtained on UCF-101, followed by HMDB-51, while performance remains lowest on SSV2. This pattern is reasonable: UCF-101 and HMDB-51 contain numerous classes that remain distinguishable and somewhat consistent even when analyzed using coarse motion cues. In contrast, actions in SSV2 are more random, involving different objects, leading to a wide variety of motion patterns within each class that our model is unable to capture due to limited training data. Compared to CLIP-based methods, our model remains clearly below the state of the art, with a gap of roughly 13–20 absolute points on HMDB-51 and UCF-101, and an even larger gap on SSV2. A direct comparison with CLIP-based methods, however, is difficult to interpret, as those models benefit from large-scale RGB-language pretraining that is absent in our setting. A more informative comparison is to R(2+1)D fine-tuned on the same splits under the same protocol: our motion-only I3D models exceed this RGB baseline, suggesting that the performance gap to CLIP-based methods is attributable to the training strategy rather than to motion representations being inherently less informative.

Taken together, the zero-shot and few-shot results show a clear trade-off. Motion-focused representations cannot match CLIP-based methods that benefit from large-scale appearance–language pretraining, but they outperform a comparable RGB baseline fine-tuned under the same protocol. This suggests that motion representations retain sufficient action-relevant information for cross-dataset transfer; the remaining gap to state-of-the-art methods reflects differences in pretraining scale and supervision rather than a fundamental limitation of motion as a signal.

4.3 Domain adaptation with privacy

Domain adaptation extends the transfer setting by incorporating unlabeled target data during training alongside labeled source data, enabling the model to explicitly reduce domain shift across datasets. Evaluation is performed on the 12-class TP-HMDB↔TP-UCF benchmark introduced by Xia et al. (2025), in both transfer directions. TP-HMDB and TP-UCF are derived from HMDB-51 and UCF-101 by retaining only the 12 overlapping action classes and remapping them to a shared label space, with privacy annotations provided by STPrivacy (Li et al., 2023b).

We compare against several baselines spanning different points on the privacy-utility spectrum. ResNet-50 serves as a frame-level appearance-based reference model for privacy attribute prediction (He et al., 2016). R(2+1)D provides a strong

TABLE 2 Few-shot action recognition top-1 accuracy (%) on HMDB-51, UCF-101, and SSV2 for $K \in \{8, 16\}$.

	HMDB-51		UCF-101		SSV2	
Method	K=8	K=16	K=8	K=16	K=8	K=16
ViFi-CLIP (Rasheed et al., 2023)	64.5	66.8	90.0	92.7	8.6	11.0
MAXI (Lin et al., 2023)	65.0	66.5	92.4	93.5	9.3	12.4
TC-CLIP (P) (Kim et al., 2024)	71.4	73.0	96.6	97.3	12.1	15.2
R(2+1)-D (RGB)	40.30	49.70	71.72	81.28	2.87	4.58
I3D OF (ours)	48.11	50.23	77.63	81.13	5.39	6.61
I3D OF + MHI (ours)	49.26	52.28	79.85	83.67	4.17	5.73

Our results are mean top-1 accuracy across validation splits reused from TC-CLIP (Kim et al., 2024); prior work numbers are reported as top-1 values from the original papers / official repositories. R(2+1)D RGB serves as a Kinetics-pretrained RGB baseline fine-tuned under the same protocol; our motion-only models exceed its accuracy despite operating without appearance information.

TABLE 3 Cross-domain results on the TP-HMDB↔TP-UCF 12-class benchmark.

	TP-HMDB → TP-UCF			TP-UCF → TP-HMDB		
Method	Top-1 ↑	cMAP ↓	F1 ↓	Top-1 ↑	cMAP ↓	F1 ↓
Source Only (RGB) (Xia et al., 2025)	85.81	72.51	0.592	78.69	68.47	0.552
VITA (Wu et al., 2022)	65.50	70.60	0.587	72.73	64.25	0.469
SPAct (Dave et al., 2022)	72.22	**66.47**	0.570	74.26	64.89	0.525
VITA+DANN (Wu et al., 2022)	67.37	70.65	**0.517**	71.81	**63.32**	**0.475**
SPAct+DANN (Dave et al., 2022)	63.39	67.87	0.563	65.25	64.44	0.517
GenPriv (Xia et al., 2025)	**87.91**	67.42	0.519	**80.55**	64.84	0.527
ResNet-50 (RGB)	-	70.09	0.675	-	66.80	0.628
R(2+1)-D + DANN (RGB)	85.47	-	-	76.20	-	-
ResNet-50 (OF)	-	69.34	0.546	-	61.43	0.493
ResNet-50 (MHI)	-	70.19	0.603	-	64.39	0.567
I3D OF + DANN	91.60	**67.98**	**0.451**	82.84	**61.01**	**0.476**
I3D OF + MHI + DANN	**93.36**	68.97	0.502	**83.89**	64.05	0.521

Results above the midline are taken from GenPriv (Xia et al., 2025). Bold indicates the best result within each half of the table. +DANN denotes domain-adversarial training adapted from Ganin et al. (2016). Our results in the bottom half are averaged over the three official UCF and HMDB train/test splits. For our I3D models, we train the domain-adaptation and privacy-attribute-prediction stages separately, using our pretrained motion model. ResNet-50 serves as a frame-level baseline for privacy prediction. Optical flow with MHI provides the best accuracy, while optical flow alone reduces sensitive attribute leakage the most.

convolutional RGB action recognition baseline (Tran et al., 2018). VITA (Wu et al., 2022), SPAct (Dave et al., 2022), and GenPriv (Xia et al., 2025) are dedicated privacy-preserving methods that apply learned transformations to RGB to suppress sensitive information. As no publicly available code was found for these methods, their results are taken directly from Xia et al. (2025). Our motion-only models are initialized from a Kinetics-pretrained checkpoint. The R(2+1)D baseline we compare against uses publicly available Kinetics-pretrained weights. Since the motion input representations are learning-free and fixed, no privacy transformation needs to be trained.

Domain adaptation follows a DANN-style objective (Ganin et al., 2016): labeled source videos supervise action recognition, unlabeled target videos are aligned through a gradient-reversal domain classifier, and a target-entropy term sharpens target predictions. Both class logits and class names are used via our dual-head setup.

Privacy leakage is measured in a second stage, where the motion model is re-initialized and adapted with a multi-attribute predictor for the five STPrivacy attributes: *face*, *skin_color*, *gender*, *nudity*, and *relationship*. Each attribute is treated as a binary classification task, predicting whether it is identifiable. The privacy attribute predictor is trained for 50 epochs. We report cMAP and F1 with respect to the positive class only, and no class-balancing loss is applied. For attribute prediction, a single frame per video is sampled at inference time, consistent with the frame-level evaluation used in prior work. These results should be interpreted with caution, as TP-HMDB↔TP-UCF exhibits both strong attribute imbalance and substantial correlation between action class and privacy attributes.

Table 3 reports our results alongside prior numbers from Xia et al. (2025). Our motion-only models exceed all prior methods in action recognition accuracy on both transfer directions, with I3D OF + MHI + DANN achieving the highest Top-1 accuracy overall. On privacy metrics, I3D OF + DANN achieves the lowest F1 scores in both directions, outperforming all methods above and below the midline.

We caution against interpreting raw F1 as direct evidence of privacy preservation, because TP-HMDB↔TP-UCF exhibits

TABLE 4 Privacy-attribute predictability relative to class-prior and action-only baselines on TP-HMDB↔TP-UCF.

Direction	Method	F1	Δ Maj. F1	Δ Action F1
TP-HMDB → TP-UCF	ResNet-50 RGB	0.675	+0.244	+0.087
	ResNet-50 OF	0.546	+0.115	-0.042
	I3D OF + DANN	**0.451**	**+0.020**	**-0.137**
	I3D OF + MHI + DANN	0.502	+0.071	-0.086
TP-UCF → TP-HMDB	ResNet-50 RGB	0.628	+0.207	+0.102
	ResNet-50 OF	0.493	+0.071	-0.034
	I3D OF + DANN	**0.475**	**+0.054**	**-0.051**
	I3D OF + MHI + DANN	0.521	+0.099	-0.006

Δ Majority F1 measures the gap above an always-majority-class predictor. Δ Action-only F1 measures the gap above a predictor that uses only the action label, quantifying how much attribute prediction exceeds what can be explained by action–attribute correlation alone. Lower values indicate less additional sensitive-attribute information. Bold indicates the best result.

strong attribute imbalance and substantial correlation between action class and privacy attributes. Table 4 therefore reports sensitive attribute predictability relative to two reference baselines: an always-majority-class predictor, which captures class-prior effects, and an action-only predictor, which captures how much attribute prediction can be explained by correlation with the action label alone. Under both baselines, I3D OF + DANN shows the strongest reduction in both transfer directions. Notably, its Δ Action F1 is negative in both directions, meaning the motion representation exposes less attribute information than the action label alone would imply. ResNet-50 on RGB shows a positive Δ Action F1, indicating that RGB carries additional sensitive information beyond what is correlated with the action class. The reduction in sensitive attribute predictability relative to our RGB baseline is comparable to that achieved by dedicated privacy-preserving methods, without requiring a learned privacy transformation or any privacy-label supervision.

To better isolate the role of language supervision, Table 5 compares domain adaptation performance when training uses class logits only vs. both text labels and class logits. From the results, we can see that text supervision yields consistent improvements across both transfer directions and motion backbones, with the largest gains observed for I3D OF + MHI. This finding indicates that text can provide a useful auxiliary signal in a supervised cross-domain setting.

TABLE 5 Text supervision comparison using split 1 of the UCF-HMDB12 dataset.

Backbone	Training signal	TP-HMDB → TP-UCF	TP-UCF → TP-HMDB
I3D OF	Class logits only	88.04	80.34
	Text + class logits	**91.55** (+3.51)	**81.46** (+1.12)
I3D OF + MHI	Class logits only	88.04	77.53
	Text + class logits	**92.78** (+4.74)	**82.87** (+5.34)

We compare models trained only with class logits against those trained with both text labels and class logits. Values are action top-1 accuracy (%), with gains from adding text shown in parentheses relative to the class-logits-only counterpart. The usage of text labels increases domain adaptation accuracy by 1–5%. Bold indicates the best result.

TABLE 6 Privacy attributes in PA-HMDB51 and their label definitions.

Attribute	Values
Skin color	0: Unidentifiable; 1: White; 2: Brown/Yellow; 3: Black; 4: Multiple
Face	0: Invisible (<10%); 1: Partial (10–70%); 2: Full (>70%)
Gender	0: Unidentifiable; 1: Male; 2: Female; 3: Multiple
Nudity	0: None (long sleeves/pants); 1: Partial (short sleeves/shorts); 2: Semi (half-naked)
Relationship	0: Unidentifiable; 1: Identifiable

4.4 Privacy-attribute prediction on PA-HMDB51

We further evaluate our model on PA-HMDB51 (Wu et al., 2022), which provides 515 privacy annotations on videos from HMDB-51. Unlike the TP-HMDB↔TP-UCF benchmark, this dataset contains nonbinary labels per class, as can be seen in Table 6. As an RGB baseline, a Vision Transformer (ViT-S) (Touvron et al., 2021) is fine-tuned on individual frames, providing a strong appearance-based reference for how much sensitive attribute information is accessible from RGB alone. We sample multiple different frames per video during training, while averaging predictions at test time to obtain a video-level score. To obtain a worst-case estimate of privacy leakage, the best-performing checkpoint across all epochs is selected based on test set performance. Each model sees only one frame per prediction. Due to the severe label imbalance in this small dataset (see Figure 6 in the PA-HMDB51 paper Wu et al., 2022), we use a class-balancing loss.

Table 7 shows that both motion representations reduce sensitive attribute predictability relative to RGB, with optical flow providing the strongest suppression overall. Skin-color prediction approaches chance level for optical flow, while face and gender prediction also remain well below RGB. MHI reduces attribute predictability relative to RGB but retains more appearance-related information than optical flow, consistent with its retention of coarse shape and silhouette cues. The ViT-S scores on motion inputs should be interpreted with caution: a ViT pretrained on RGB is not well adapted to optical flow or MHI, and its near-chance scores likely reflect an inability to extract meaningful features from unfamiliar input formats rather than genuine privacy

TABLE 7 PA-HMDB51 privacy-attribute prediction results reported as macro F1 (mean ± std over 3 folds), where lower values indicate stronger privacy preservation.

Method	Gender	Skin color	Face	Nudity	Relationship
Label-distribution baselines					
Majority class	0.19	0.16	0.20	0.21	0.46
Action label	0.38	0.16	0.52	0.41	0.66
Appearance baseline					
ViT-S RGB	0.38 ± 0.04	0.36 ± 0.05	0.62 ± 0.03	0.64 ± 0.03	0.60 ± 0.03
Motion representations					
ViT-S OF only	**0.07 ± 0.03**	**0.08 ± 0.03**	**0.40 ± 0.04**	**0.33 ± 0.04**	**0.56 ± 0.08**
ViT-S MHI only	0.28 ± 0.04	0.32 ± 0.04	0.52 ± 0.04	0.50 ± 0.04	0.65 ± 0.04
I3D OF only	0.33 ± 0.05	0.20 ± 0.09	0.47 ± 0.01	0.45 ± 0.02	0.66 ± 0.04
I3D MHI + OF	0.36 ± 0.04	0.27 ± 0.07	0.52 ± 0.02	0.45 ± 0.02	0.61 ± 0.04

PA-HMDB51 has severely imbalanced labels, so F1 scores can appear non-trivially high without any genuine attribute prediction: the majority-class baseline captures the contribution of label skew alone, while the action-only baseline captures how much attribute leakage is already implicit in knowing the action category. Both motion representations reduce the predictability of sensitive attributes relative to RGB, with dense Farnebäck optical flow providing the strongest overall appearance suppression. Bold indicates the best result.

suppression. The I3D results are more informative, as those models are pretrained on motion and have learned to extract action-relevant information from it—the low attribute prediction scores therefore provide stronger evidence that optical flow genuinely suppresses sensitive appearance information.

4.5 Ablation studies

We perform a series of controlled ablation studies to isolate the effect of the main design choices in our pipeline on few-shot transfer from UCF-101 to HMDB-51 over the overlapping 12 classes ($C = 12$) in the $K = 16$-shot setting. Unless otherwise stated, each experiment modifies a single factor while keeping the remaining training and evaluation settings fixed, and results are reported as mean top-1 accuracy ± standard deviation across at least two random seeds.

We study four complementary aspects of the method, progressing from input design to higher-level modeling choices. Table 8 first examines motion preprocessing, including the frame-difference threshold, the number of MHI windows, and the Farneback optical-flow preset, followed by Table 9, which evaluates the role of input shape and temporal support through image resolution, the number of MHI frames, the number of flow frames, and flow resolution. Next, Table 10 studies the effect of text supervision by contrasting different text banks and text-conditioning strategies. Finally, Table 11 evaluates the impact of representation mixing as a higher-level modeling component.

4.6 Computational efficiency

Table 12 reports model size, computational cost, and inference throughput for our motion models alongside the RGB baselines. All measurements use batch size 1, single-threaded CPU or single GPU, and exclude preprocessing. Preprocessing videos to generate Farnebäck dense optical flow and MHI adds little overhead and can be parallelized across CPU cores.

TABLE 8 Motion preprocessing ablation on UCF12→HMDB12.

Threshold	Windows	FB preset	UCF val motion	HMDB motion
10	15	Default	43.85 ± 6.56	28.47 ± 5.86
15	10	Default	40.27 ± 7.34	26.56 ± 5.44
15	15	Default	41.65 ± 3.82	27.08 ± 1.38
15	15	Smooth	47.70 ± 3.63	27.08 ± 1.38
15	20	Default	55.33 ± 3.63	30.38 ± 1.20
15	25	Default	**56.43±6.06**	**33.16±1.68**
20	15	Default	47.08 ± 0.43	30.38 ± 6.08
25	15	Default	45.22 ± 7.45	25.18 ± 3.18

All runs use the same training protocol and differ only in frame-difference threshold, MHI window count, and Farneback preset. Results are top-1 accuracy (%, mean ± std over 3 seeds). Bold indicates the best result.

I3D OF requires 93 GFLOPs per clip and runs at 5.3 ms per video on GPU, comparable to R(2+1)D (162 GFLOPs, 5.4 ms) while using approximately half the parameters and 25% less GPU memory. I3D MHI+OF doubles the parameter count relative to I3D OF, as expected for a two-stream model, but remains well below R(2+1)D in GFLOPs and substantially below TC-CLIP in every dimension. TC-CLIP requires 4.9× more parameters, 5.6× more GFLOPs and 6.8× more GPU memory than I3D MHI+OF, with GPU inference 10× slower, which is relevant for multi-camera deployments where throughput and memory per stream determine whether inference is feasible on shared hardware.

On CPU, the gap is larger. TC-CLIP requires nearly 10 s per video single-threaded, making multi-camera CPU deployment impractical. I3D MHI+OF requires 3.4 s, which is still too slow for real-time use on CPU. A lightweight alternative based on X3D-XS is discussed in Supplementary Section 5; that model runs at ∼300 ms per video on CPU with only 3.3 GFLOPs, at a modest cost in accuracy.

TABLE 9 Shape and temporal-support ablation trained for 10 epochs, all other parameters are fixed.

Image size	MHI frames	Flow frames	Flow resolution	UCF val motion	HMDB motion
224	32	128	112	41.65 ± 3.82	**27.08 ± 1.38**
224	32	64	112	41.23 ± 1.35	25.70 ± 3.01
224	32	32	224	45.43 ± 4.43	25.87 ± 6.55
224	16	64	224	**48.87 ± 4.47**	26.74 ± 5.57
112	128	32	224	29.00 ± 1.96	22.05 ± 1.83

Bold indicates the best result. We chose the first configuration as we focus on cross-domain results.

TABLE 10 Ablation on the benefit of text supervision, trained for 20 epochs.

Text bank	Text mode	UCF12 val	HMDB12	HMDB16
Descriptions	`Averaged`	66.60 ± 2.70	**40.45 ± 5.74**	31.77 ± 6.35
Descriptions	`Multipos`	**69.76 ± 2.49**	38.37 ± 4.05	**33.98 ± 2.38**
Labels	`Labels`	68.38 ± 2.20	38.89 ± 7.55	31.64 ± 5.83

Bold indicates the best result. We chose the multiple positive supervision, where all descriptions belonging to the same class are considered correct in the loss.

5 Discussion

Motion-centric action recognition introduces a practical trade-off. Compared with RGB-based pipelines, it requires additional preprocessing and offers fewer pretrained backbones. In return, it imposes a more controlled inductive bias: by suppressing most static appearance, the model is pushed to rely more on temporal structure than on scene, identity, or object-context shortcuts. This does not remove all non-action cues. In datasets such as Kinetics, camera motion and viewpoint can still encode coarse scene information, and factors such as camera angle or subject scale may remain exploitable. Motion-focused learning is therefore not immune to shortcuts, but it reduces the set of shortcuts available to the model.

The privacy properties of the two motion representations also differ. Optical flow suppresses appearance most strongly, whereas MHI can retain residual structure when subjects are close to the camera and moving. With a grayscale threshold of $\delta = 15$, weak pixel variation is removed and fine detail is limited, but coarse contours and some clothing patterns may still remain visible. In such cases, broad body shape or a rough male–female distinction may still be inferable, even if finer attributes are much less discernible.

A second trade-off concerns CLIP-based text supervision. Its main advantage is semantic structure: related labels can help organize the embedding space and support transfer across datasets. However, this same structure can be counterproductive when textual similarity does not reflect motion similarity, especially for fine-grained actions distinguished primarily by dynamics. This limitation is likely amplified in our setting because the model is trained only on Kinetics-400. Consider, for example, the classes *Playing Guitar* and *Playing Sitar*: these actions share nearly identical motion patterns, providing insufficient signal for the model to distinguish between them. Supplementary Figure S3 illustrates this confusion by comparing the per-class scores of our motion model with those of CLIP. Multiple alternative approaches were explored, but none yielded consistent improvements (see Supplementary Section 3). The current model with text descriptions, representation mixing, and an additional classifier head barely improves over our earliest experiment, which used only cross-entropy after dot-product of the video embeddings with the CLIP embeddings. The training strategy, the model architecture, and the pretraining dataset likely all contributed to this low accuracy. A larger-scale motion–text pretraining setup, such as How-to-100M (Miech et al., 2019), could improve alignment and strengthen zero-shot performance, but this remains an open question.

TABLE 11 Representation mixing experiment, trained for 20 epochs, all other parameters are fixed.

Representation mixing	UCF12 val	HMDB12	HMDB16
on	69.07 ± 3.75	**42.88 ± 1.08**	**36.07 ± 0.60**
off	**70.03 ± 2.78**	41.15 ± 2.90	34.38 ± 3.41

Bold indicates the best result. We selected representation mixing because it provides a larger set of label combinations, thereby offering greater flexibility.

On a brighter note, the training framework is not limited to RGB-derived motion, allowing for its extension to other camera inputs. Since both optical flow and MHI are driven by temporal change rather than semantic appearance, the same principle can be applied to privacy-friendly sensors like depth, infrared, and event cameras. Depth and infrared are the most direct extensions, since frame differences and motion fields can be computed with only minor changes to the pipeline. In scene-depth video, these cues would capture geometric change while discarding color and texture almost entirely, potentially reducing identity leakage even further. Infrared likewise preserves temporal activity under low illumination while suppressing much of the fine appearance detail present in RGB, though noisy infrared sensors, such as those in the Microsoft Kinect V2, would likely require additional preprocessing.

Event cameras are less straightforward because they output asynchronous brightness-change events rather than frames. Nevertheless, short-window accumulation should enable the derivation of a usable motion representation. Let an event stream be represented as $e_i = (x_i, y_i, t_i, p_i)$, where (x_i, y_i) denotes pixel location, t_i the timestamp, and $p_i \in \{-1, +1\}$ the polarity. A

TABLE 12 Computational profile of motion and RGB models at batch size 1.

Model	Params (M)	GFLOPs	GPU (ms/vid)	CPU (ms/vid)	Peak GPU mem (MB)
I3D OF (ours)	13.5	93.0	5.4	2487	1575
I3D MHI+OF (ours)	26.3	139.9	10.2	3370	2384
R(2+1)D (RGB)	31.5	162.6	5.5	3461	2095
TC-CLIP (RGB)	127.5	781.0	105.0	9932	16245

GPU timings use an NVIDIA RTX Pro 6000 with AMP; CPU timings are single-threaded.

frame-like signal can then be constructed over a temporal bin $[t, t + \Delta t)$ by accumulating events. The resulting pseudo-frames are suitable for MHI updates and, in principle, for approximate flow estimation. These extensions suggest that the core idea of motion-focused, appearance-suppressed action recognition may generalize beyond conventional video.

We also see clear potential in synthetic data as a future direction. In principle, synthetic action datasets could enable systematic control over viewpoint, lighting, background, body shape, clothing, and demographic balance far more than is feasible with real-world collections. This control and customizability would be highly valuable for studying cross-domain generalization and bias. We would have liked to include experiments with resources such as SURREACT, introduced as a synthetic action-recognition benchmark for improving viewpoint robustness (Varol et al., 2021), and BABEL, which provides language labels for large-scale motion-capture data (Punnakkal et al., 2021). In practice, however, these sources are less suitable for our setting: SURREACT remains limited by the visual realism of rendered animations, while BABEL is centered on mocap sequences and linguistic annotations rather than realistic video appearance, making transfer to real cross-domain video recognition difficult to interpret. More broadly, current synthetic resources still tend to fall short in either animation fidelity, scene realism, or label consistency. This is unfortunate, because higher-quality synthetic action data could become a powerful route toward more controlled and potentially less biased benchmarks, where sensitive factors can be balanced by design instead of only corrected after collection.

Another approach is to use generative video models; if the realism and temporal consistency of generated video continue to improve, generative synthesis may substantially reduce reliance on real-world data collection in privacy-sensitive and cross-domain settings.

6 Conclusion

We studied whether action recognition remains useful across domains when appearance is removed and only learning-free motion representations are used. Our results demonstrate that this is indeed the case, but with a clear trade-off. In domain adaptation, motion-only models match or exceed prior privacy-preserving methods in recognition accuracy, demonstrating that motion representations carry sufficient action-relevant information for cross-domain transfer under domain alignment. Zero-shot and few-shot accuracy remain below strong RGB- and CLIP-based baselines, a gap we attribute primarily to the absence of large-scale RGB–language pretraining rather than a fundamental limitation of motion as a signal. Among the motion representations evaluated, OF+MHI achieves higher recognition accuracy than optical flow alone in most settings.

Our second question concerned the degree to which motion-focused inputs suppress sensitivity to appearance-related attributes. The results provide a clear answer: motion representations consistently retain less information for predicting sensitive attributes than RGB representations, with optical flow providing the strongest suppression across all tested inputs. On PA-HMDB51, optical flow approaches chance level for skin-color prediction and remains below RGB across most attributes; the same pattern holds in the TP-HMDB↔TP-UCF experiments. The choice between optical flow and optical flow with MHI therefore depends on the deployment context: adding MHI is preferable when recognition accuracy is the priority, while optical flow is the stronger choice when minimizing sensitive attribute leakage is paramount.

Taken together, these results show that learning-free motion representations offer a practical privacy–utility trade-off for cross-domain action recognition, without requiring any learned privacy transformation. The motion pipeline is also computationally efficient: I3D MHI+OF requires $\sim$ 4.9× fewer parameters and $\sim$ 6.8× less GPU memory than TC-CLIP, and a lightweight X3D-based variant runs at $\sim$ 330 ms per video on CPU (Supplementary Section 5), making the approach feasible on shared or edge hardware.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material. The code for the conversion of the publicly available datasets used in this research is available on https://github.com/pascalbenschopTU/appearance_free_cross_domain_action_recognition. Further inquiries can be directed to the corresponding author.

Author contributions

PB: Validation, Methodology, Visualization, Formal analysis, Data curation, Software, Conceptualization, Investigation, Writing – review & editing, Writing – original draft. JG: Supervision, Writing – review & editing. JM: Supervision, Writing – review & editing. JD: Writing – review & editing, Supervision.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was supported by the Model-Driven Decisions Lab (MoDDL), a collaboration between TU Delft and the Dutch National Police.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. We used AI-based tools and large language models to support parts of the research and writing process, including language editing, code drafting assistance, brainstorming, and assistance with organizing or inspecting experimental logs for subsequent manual analysis. All generated text, code, and suggestions were reviewed and validated by the author(s) before use. The author(s) retained full responsibility for the design of the study, execution and verification of experiments, interpretation of results, and the final content of the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimag.2026.1846329/full#supplementary-material

References

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., et al. (2017). A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7388–7397. doi: 10.1109/CVPR.2017.781

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). "ViViT: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 6836–6846.

Batchuluun, G., Nguyen, D. T., Pham, T. D., Park, C., and Park, K. R. (2019). Action recognition from thermal videos. *IEEE Access* 7, 103893–103917. doi: 10.1109/ACCESS.2019.2931804

Bertasius, G., Wang, H., and Torresani, L. (2021). "Is space-time attention all you need for video understanding?" in *Proceedings of the International Conference on Machine Learning (ICML)*, eds. M. Marina and . Tong (PMLR).

Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., and Andreopoulos, Y. (2020). Graph-based spatio temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing* 29, 9084–9098. doi: 10.1109/TIP.2020.3023597

Bobick, A. F., and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 257–267. doi: 10.1109/34.910878

Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., and Chalupka, K. (2020). "Rethinking zero-shot video classification: End-to-end training for realistic applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 4613–4623.

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA: IEEE Computer Society), 4724–4733. doi: 10.1109/CVPR.2017.502

Chi, H.-G., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., and Ramani, K. (2022). "InfoGCN: Representation learning for human skeleton-based action recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA: IEEE), 20154–20164.

Dave, I. R., Chen, C., and Shah, M. (2022). "Spact: Self-supervised privacy preservation for action recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA: IEEE), 20132–20141.

Doshi, K., and Ylmaz, Y. (2023). "Zero-shot action recognition with transformer-based video semantic embedding," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Vancouver, BC: IEEE), 4859–4868.

Fan, C., and Chowdhury, T. (2025). "When pose estimation fails: Measuring occlusion for reliable multimodal interaction," in *Companion Proceedings of the 27th International Conference on Multimodal Interaction, ICMI Companion '25* (New York, NY: Association for Computing Machinery), 58–64.

Farnebäck, G. (2003). "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, eds. J. Bigun, and T. Gustavsson (Berlin: Springer Berlin Heidelberg), 363–370.

Feichtenhofer, C. (2020). "X3D: expanding architectures for efficient video recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Piscataway, NJ), 200–210. doi: 10.1109/CVPR42600.2020.00028

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 1–35. doi: 10.5555/2946645.2946704

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., et al. (2017). "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision* (Piscataway, NJ), 5842–5850. doi: 10.1109/ICCV.2017.622

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ), 770–778. doi: 10.1109/CVPR.2016.90

Ilic, F., Pock, T., and Wildes, R. P. (2022). "Is appearance free action recognition possible?" in *European Conference on Computer Vision (ECCV)* (Berlin; Heidelberg: Springer-Verlag). doi: 10.1007/978-3-031-19772-7_10

Innocenti, S. U., Becattini, F., Pernici, F., and Del Bimbo, A. (2021). Temporal binary representation for event-based action recognition. In *25th International Conference on Pattern Recognition* (ICPR). 10426–10432. doi: 10.1109/ICPR48806.2021.9412991

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., et al. (2017). The kinetics human action video dataset. *arXiv* [preprint] arXiv:1705.06950. doi: 10.48550/arXiv.1705.06950

Kim, M., Han, D., Kim, T., and Han, B. (2024). "Leveraging temporal contextualization for video action recognition," in *European Conference on Computer Vision (ECCV)* (Berlin; Heidelberg: Springer-Verlag).

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). *Hmdb51: A Large Video Database for Human Motion Recognition*, 2556–2563.

Kumawat, S., and Nagahara, H. (2022). "Privacy-preserving action recognition via motion difference quantization," in *European Conference on Computer Vision* (Cham: Springer), 518–534.

Li, H., Liu, Y., Zhang, H., and Li, B. (2023a). "Mitigating and evaluating static bias of action representations in the background and the foreground," in *International Conference on Computer Vision (ICCV)*.

Li, M., Xu, X., Fan, H., Zhou, P., Liu, J., Liu, J.-W., et al. (2023b). "Stprivacy: Spatio-temporal privacy-preserving action recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Los Alamitos, CA: IEEE Computer Society), 5083–5092. doi: 10.1109/ICCV51070.2023.01823

Lin, W., Karlinsky, L., Shvetsova, N., Possegger, H., Kozinski, M., Panda, R., et al. (2023). "Match, expand and improve: unsupervised finetuning for zero-shot action recognition with language knowledge," in *ICCV* (Piscataway, NJ: IEEE). doi: 10.1109/ICCV51070.2023.00267

Luo, R., Anand, A., Zulkernine, F., and Rivest, F. (2024). YOLO-i3D: Optimizing inflated 3d models for real-time human activity recognition. *J. Imag.* 10:269. doi: 10.3390/jimaging10110269

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Piscataway, NJ: IEEE), 2630–2640.

Osokin, D. (2018). Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv* [preprint] arXiv:1811.12004. doi: 10.5220/0007555407440748

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., et al. (2022). "Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22* (New York, NY: Association for Computing Machinery), 2302–2314.

Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., and Black, M. J. (2021). "Babel: Bodies, action and behavior with english labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 722–731.

Rasheed, H., Khattak, M. U., Maaz, M., Khan, S., and Khan, F. S. (2023). "Finetuned clip models are efficient video learners," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society).

Ren, B., Liu, M., Ding, R., and Liu, H. (2024). A survey on 3D skeleton-based action recognition using learning method. *Comp. Biol. Syst*. doi: 10.34133/cbsystems.0100

Sánchez-Caballero, A., de López-Diz, S., Fuentes-Jiménez, D., Losada-Gutiérrez, C., Marrón-Romera, M., Casillas-Perez, D., et al. (2022). 3DFCNN: Real-time action recognition using 3D deep neural networks with raw depth information. *Multimedia Tools and Applications* 81, 24119–24143. doi: 10.1007/s11042-022-12091-z

Sevilla-Lara, L., Liao, Y., Güney, F., Jampani, V., Geiger, A., and Black, M. J. (2017). "On the integration of optical flow and action recognition," in *German Conference on Pattern Recognition*.

Shou, Z., Lin, X., Kalantidis, Y., Sevilla-Lara, L., Rohrbach, M., Chang, S.-F., et al. (2019). "DMC-Net: Generating discriminative motion cues for fast compressed video action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1268–1277.

Simonyan, K., and Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14* (Cambridge, MA: MIT Press), 568–576.

Soomro, K., Zamir, A., and Shah, M. (2012). UCF101: a dataset of 101 human actions classes from videos in the wild. *arXiv* [preprint] arXiv.1212.0402. doi: 10.48550/arXiv.1212.0402

Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. *CoRR*, abs/1801.04264. doi: 10.1109/CVPR.2018.00678

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila, and T. Zhang (New York: Proceedings of Machine Learning Research), 10347–10357.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6450–6459.

Varol, G., Laptev, I., Schmid, C., and Zisserman, A. (2021). "Synthetic humans for action recognition from unseen viewpoints," in *IJCV* (Springer Nature). doi: 10.1007/s11263-021-01467-7

Wang, C., Zhang, F., Zhu, X., and Ge, S. S. (2021). *Low-Resolution Human Pose Estimation*.

Wang, H., and Schmid, C. (2013). "Action recognition with improved trajectories," in *2013 IEEE International Conference on Computer Vision*, 3551–3558.

Wang, P., Li, W., Ogunbona, P., Wan, J., and Escalera, S. (2018). RGB-D-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding* 171, 118–139. doi: 10.1016/j.cviu.2018.04.007

Wei, L., Yu, X., and Liu, Z. (2024). Human pose estimation in crowded scenes using keypoint likelihood variance reduction. *Displays*. 102675. doi: 10.1016/j.displa.2024.102675

Wei, P., Kong, L., Qu, X., Ren, Y., Xu, Z., Jiang, J., et al. (2023). "Unsupervised video domain adaptation for action recognition: a disentanglement perspective," in *Thirty-Seventh Conference on Neural Information Processing Systems* (Cham: Springer International Publishing).

Wu, C.-Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A. J., and Krähenbühl, P. (2018). "Compressed video action recognition," in *CVPR* (Los Alamitos, CA: IEEE Computer Society).

Wu, Z., Wang, H., Wang, Z., Jin, H., and Wang, Z. (2022). Privacy-preserving deep action recognition: an adversarial learning framework and a new dataset. *IEEE Trans. Pattern Analy. Mach. Intellig.* 44, 2126–2139. doi: 10.1109/TPAMI.2020.3026709

Xia, Z.-W., Lin, K.-Y., Li, Y.-M., Huang, W.-J., Tan, X.-T., and Zheng, W.-S. (2025). "Less static, more private: towards transferable privacy-preserving action recognition by generative decoupled learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Piscataway, NJ: IEEE), 12894–12903. doi: 10.5555/2946645.2946704

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18* (New York: AAAI Press).

Zhai, Y., Liu, Z., Wu, Z., Wu, Y., Zhou, C., Doermann, D., et al. (2023). "Soar: Scene-debiasing open-set action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Piscataway, NJ: IEEE), 10244–10254.