ACTIONS SPEAK LOUDER THAN FRAMES: QUANTIFYING STATIC AND TEMPORAL RELIANCE IN ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video action recognition models can base their predictions on two qualitatively different sources of evidence: the visual appearance present within individual frames and the temporal dynamics that unfold across frames. Existing explanation methods typically conflate these sources within a single spatiotemporal attribution, making it difficult to determine and quantify how much of the prediction is supported by static appearance, motion, or both. To distinguish genuine temporal evidence from confidence already recoverable from static frames, we introduce **Motion Headroom**, which also provides a criterion for when temporal attribution is identifiable. We then introduce **Reliance Probes** that measure how removing appearance or temporal variation affects noun and verb predictions, and use these probes to define a **Static-Bias Index (SBI)** for auditing and quantifying static reliance in trained video recognition models. To localize this reliance, we use FAM (Factored Appearance and Motion), a post-hoc framework that independently optimizes appearance and motion explanations.Across EGTEA Gaze+ and Something-Something V2, the audit reveals greater static reliance on EGTEA and stronger temporal reliance on Something-Something V2 across multiple architectures. We further define an explanation-conditioned SBI that measures how much of the model's temporal reliance is captured by the temporal evidence selected by FAM. Finally, we introduce an axis-specific evaluation protocol that evaluates appearance and motion explanations against the prediction component each axis is intended to explain.

1 INTRODUCTION

Video action recognition is inherently temporal, yet a correct prediction need not depend strongly on temporal evidence. A model may recognize an action from objects, scene context, or other appearance cues visible in individual frames, or from changes that unfold across the video. These two sources of evidence have different implications for interpreting model behavior that is not captured by top-1 accuracy on a held-out test split. Distinguishing them is therefore important for understanding what action-recognition models actually use when making a prediction and to what extent. This question applies across modern architectures such as TSM (Lin et al., 2019), SlowFast (Feichtenhofer et al., 2019), Timesformer (Bertasius et al., 2021) and Video Swin (Liu et al., 2022), particularly when such models are used in downstream tasks such as assistive robotics (Damen et al., 2018) and surgical workflow analysis (Valderrama et al., 2022), where the basis of a prediction matters.

Existing explanation methods do not make this distinction explicit. For static images, gradient and perturbation based methods (Selvaraju et al., 2017; Wang et al., 2020; Petsiuk et al., 2018; Fong et al., 2019) identify spatial regions associated with a prediction. Their video extensions typically produce a single spatiotemporal attribution (Fong et al., 2019; Li et al., 2021), combining information about *where* relevant appearance occurs with *when* temporal evidence matters. As a result, the explanation does not reveal whether a prediction is supported primarily by within-frame appearance or by temporal dynamics across frames. We refer to the former as *appearance* evidence, including objects, texture, color, and within-frame interactions, and to the latter as *motion* evidence, describing how visual content changes over time.

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.

This distinction matters because action recognition models can rely heavily on frame-level appearance cues rather than combining appearance and motion information(Kowal et al., 2022; Ilic et al., 2022). If an action recognition model can recover its prediction from a single static frame, this indicates overfitting to static correlations in the training data, and such a model may still yield high benchmark accuracy while yielding brittle predictions that fail when the background, objects or appearance change. After all, the objective of action recognition is to identify an action from its temporal dynamics, rather than to infer action from static objects or scene context. Prior work measures such static bias in intermediate representations (Kowal et al., 2022),or through modified training and evaluation settings (Ilic et al., 2022; Li et al., 2023), but these analyses primarily characterize a model or dataset as a whole. They do not show whether a particular prediction depends on temporal variation, or which frames provide that dependence. This distinction is useful, for example, for separating instrument appearance from procedure-specific motion in clinical video. In egocentric assistive technology, it can distinguish active user intent, such as "reaching for X," from passive context, such as "X is present."

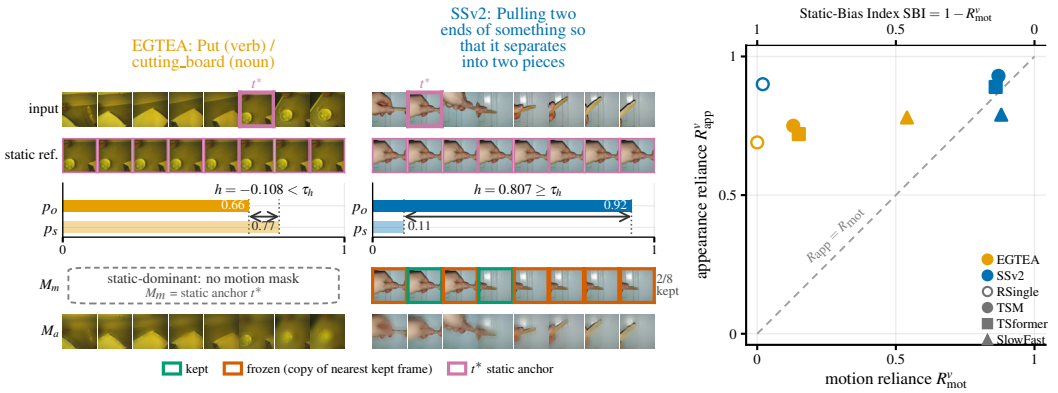
Separating appearance and motion also raises an identifiability problem. Temporal explanations are commonly optimized to retain the target confidence when selected frames are kept, but a temporal mask can appear sufficient even when a repeated static frame already provides high confidence. In this case, the temporal contribution cannot be reliably separated from static support. Temporal attribution should therefore report when it is identifiable and, when identifiable, measure an explanation against only the confidence that static appearance cannot account for. For such clips, Factored Appearance and Motion (FAM) optimizes temporal explanations against this motion-attributable confidence while separately explaining appearance evidence. Our contributions are:

1. **Motion headroom for identifiable temporal explanations.** (§3.2) We introduce *motion headroom*, which measures target confidence available beyond a static-video baseline. It provides an explicit criterion for when temporal attribution is identifiable. For identifiable clips, headroom-normalized sufficiency measures how much motion-dependent confidence is retained by the selected frames.
2. **Cross-axis appearance and motion reliance probes.** (§4) We introduce post-hoc probes that measure how reducing appearance or temporal variation affects noun and verb predictions. This four-way analysis captures both the expected appearance–noun and motion–verb relationships and the cross-axis appearance–verb and motion–noun relationships, avoiding the assumption that each prediction head depends on only one evidence type.
3. **Post-hoc diagnosis and quantification of static bias.** (§4) Using the appearance–motion factorization, we introduce a Static-Bias Index (SBI) and related diagnostics for measuring whether video predictions rely on temporal variation or can be supported by static appearance. Across TSM, TimeSformer, and SlowFast, the audit identifies greater static reliance on EGTEA and stronger temporal reliance on Something-Something V2, while a single-frame control provides the expected static-dominant behavior.
4. **Axis-specific evaluation for factored explanations.** (§5) We introduce a *separation evaluation protocol* that evaluates each axis independently. Appearance masks are scored by insertion/deletion on the noun, while motion masks are scored by headroom-normalized sufficiency and necessity on the verb, restricted to motion-identifiable clips.

We evaluate on EGTEA Gaze+ (Li et al., 2018a), and Something-Something V2 (Goyal et al., 2017) against STEP (Li et al., 2021) and Extremal Perturbation (Fong et al., 2019). We use two controlled ablations and an exhaustive exact solution when it is feasible to compute to understand how much larger our motion explanations are compared to the optimal explanation size. FAM instantiates the two explanation axes using Gaussian blur for appearance and nearest-frame replacement(Vasu et al., 2026) to perturb motion. These perturbations serve as the intervention mechanisms. The contributions of this work concern how appearance and temporal evidence are factored, when temporal attribution is identifiable, how the resulting reliance is quantified, and how the two explanation axes are evaluated rather than the interventions themselves.

2 Related work

Several works have shown that action-recognition models can rely strongly on static appearance rather than temporal dynamics. Single-frame and order-insensitive analyses show that many benchmark



(a) (b)

Figure 1: **Motion headroom and static-bias audit.** (a) Each clip is compared with its static reference, formed by repeating the most confident frame t^*. EGTEA is static-dominant ($h < \tau_h$), so FAM returns only M_a; SSv2 has identifiable temporal evidence and returns both M_m (green: kept, orange: frozen) and M_a. (b) Verb motion/appearance reliance and corresponding SBI across models and datasets. SSv2 shows strong motion reliance across architectures, while EGTEA is more appearance-reliant except for SlowFast; single-frame controls show near-zero motion reliance.

actions can be recognized without substantial temporal reasoning (Huang et al., 2018; Sevilla-Lara et al., 2021), motivating more temporally demanding datasets such as Something-Something (Goyal et al., 2017). Other work studies scene, representation, foreground, and background bias in action recognition (Li et al., 2018b; Choi et al., 2019; Kowal et al., 2022; Li et al., 2023), while Ilic et al. (2022) directly evaluates recognition after removing informative static appearance. These approaches primarily characterize bias at the model or dataset level through retraining, representation analysis, or specialized evaluation settings. In contrast, our Reliance Probes operate post-hoc on an already-trained model, measuring for each prediction how confidence changes when appearance or temporal variation is reduced. SBI aggregates these measurements to quantify static reliance, while the explanation-conditioned SBI measures how much of the model's temporal reliance is captured by a particular temporal explanation.

Existing video explanation methods generally produce a single spatiotemporal attribution, making spatial appearance and temporal importance difficult to separate. Gradient- and activation-based approaches include Temporal GradCAM, SWAG-V, and Saliency Tubes (Selvaraju et al., 2017; Hartley et al., 2022; Stergiou et al., 2019). Perturbation methods such as Extremal Perturbation (EP) (Fong et al., 2019) and STEP (Li et al., 2021) optimize spatiotemporal masks, with STEP adding a temporal smoothness prior; we use both as baselines in Section 5. FAM instead optimizes appearance and motion explanations independently. For the temporal intervention, we use the nearest-frame replacement mechanism of Vasu et al. (2026); our contribution lies in the factored objectives, the static reference used to determine when temporal attribution is identifiable, and the use of the resulting explanations to quantify model reliance.

Perturbation explanations also depend on what replaces removed evidence. Blurred references are commonly used for spatial explanations (Khorram et al., 2021; Xing et al., 2025), and the choice of reference can substantially affect the resulting attribution (Sturmfels et al., 2020). For temporal explanations, this issue is especially important because a model may retain high confidence even after temporal variation is removed. Motion Headroom explicitly measures the confidence available beyond the strongest static reference and treats low-headroom predictions as temporally unidentifiable. Standard explanation metrics such as insertion/deletion (Petsiuk et al., 2018) and sufficiency/comprehensiveness (DeYoung et al., 2020) score a single explanation against a single model output. Our separation protocol evaluates appearance and motion independently, using the prediction component associated with each axis and normalizing temporal scores by motion headroom.

Concept-based methods provide a complementary form of video interpretability. VTCD (Kowal et al., 2024), spatiotemporal TCAV extensions (Ji et al., 2023), and related analyses (Saha et al., 2024) explain predictions through higher-level concepts rather than input-level regions or frames. Language-based approaches similarly target semantic explanations (Zhuo et al., 2019). Most closely

related to our factorization, DANCE (Lee et al., 2025) retrains a concept-bottleneck model to separate pose-based motion concepts from spatial context. FAM instead operates post-hoc on frozen classifiers, requires no predefined concept vocabulary, and returns explicit spatial and temporal explanations that can also be aggregated into model-level reliance diagnostics.

3 Factored Appearance and Motion Explanations

Let f be a frozen video classifier applied to a clip $\mathbf{x} \in \mathbb{R}^{3\times T\times H\times W}$ of T uniformly sampled RGB frames at $H \times W$ resolution. For dual-head models (EGTEA), f produces verb and noun probabilities, and v^* and n^* denote the verb and noun predicted on the clean clip. For single-head models (SSv2), f predicts a single action class c^*, and we set $v^* = n^* = c^*$. FAM produces two masks,

$$M_m \in \{0,1\}^T \qquad \text{(motion: } \textit{when}\text{, binary)},$$
$$M_a \in [0,1]^{T\times H\times W} \qquad \text{(appearance: } \textit{where}\text{, soft)},$$

where M_m is optimized for the verb and M_a for the noun. The two masks are optimized separately and do not share parameters or loss terms.

3.1 Perturbation operators

Each mask is constructed using a perturbations that removes one type of evidence while keeping the clip as close as possible to the original. To remove temporal variation without introducing colors, textures or objects that are not present in the clip, we use the nearest-active-frame replacement operator of Vasu et al. (2026). Each frame with $m_t = 0$ is replaced by an unaltered copy of the closest frame in time with $m_t = 1$, which locally pauses the video. Gradients are passed through the binary mask and the frame selection using straight-through estimators. We write the motion-perturbed clip as $\hat{\mathbf{x}} = P(\mathbf{x};\, M_m)$and give the full operator in Appendix A.3. The operator is not a contribution of this paper. We use it as a fixed intervention for removing temporal variation, and test whether our findings depend on this choice in Appendices A.10 and A.12.

Replacing masked regions with a constant value such as black can introduce artificial boundaries that affect the prediction (Fong & Vedaldi, 2017; Petsiuk et al., 2018; Khorram et al., 2021). We instead blend the clip with a blurred copy $\tilde{\mathbf{x}} = \mathcal{G}_B * \mathbf{x}$, where $\mathcal{G}_B$ is a Gaussian filter with kernel size $k_B = 31$ and standard deviation $\sigma_B = 20$. Strong blur removes fine appearance evidence while preserving coarse spatial structure. The insertion and deletion clips are

$$\hat{\mathbf{x}}^{\text{ins}} = M_a \odot \mathbf{x} + (1 - M_a) \odot \tilde{\mathbf{x}}, \qquad \hat{\mathbf{x}}^{\text{del}} = (1 - M_a) \odot \mathbf{x} + M_a \odot \tilde{\mathbf{x}}, \tag{1}$$

so that $\hat{\mathbf{x}}^{\text{ins}}$ keeps only the content selected by M_a and $\hat{\mathbf{x}}^{\text{del}}$ keeps only its complement.

3.2 Motion headroom and identifiability

Video frames are highly redundant, and a single frame can often support the target prediction through the objects and scene it contains. A temporal explanation that only preserves the target confidence can therefore appear faithful without capturing any evidence that depends on motion. We separate the two by comparing each clip with a static reference.

Definition 1 (Static reference)**.** Let $p_s^t = p_{v^*}\big(\mathrm{Repeat}(\mathbf{x}_t, T)\big)$ be the target confidence when frame t is repeated across the clip. The static reference is $\mathbf{x}^{\text{stat}} = \mathrm{Repeat}(\mathbf{x}_{a^*}, T)$, where $a^* = \arg\max_t p_s^t$.

The static reference gives the highest target confidence the model assigns to the clip when all temporal variation is removed. Selecting the best frame, rather than a fixed frame such as the first or middle frame, makes the reference conservative: confidence above $p_{v^*}(\mathbf{x}^{\text{stat}})$ cannot be explained by the appearance of any single frame. $\mathbf{x}^{\text{stat}}$ is also the output of P when M_m contains no active frames.

Definition 2 (Motion headroom)**.** Let $p_o = p_{v^*}(\mathbf{x})$ and $p_s = p_{v^*}(\mathbf{x}^{\text{stat}}) = \max_t p_s^t$. The motion headroom is $h = p_o - p_s$. A clip is *motion-identifiable* if $h \geq \tau_h$, and *static-dominant* otherwise.

A large h indicates that temporal variation provides confidence beyond what any single frame provides. A small or negative h indicates that the prediction can already be supported by static appearance, so the contribution of motion cannot be reliably isolated. We use $\tau_h = 0.05$, and show in Appendix A.5 that our findings do not change for $\tau_h \in [0.01, 0.20]$.

Headroom-normalized scores. Let $p_k = p_{v^*}\big(P(\mathbf{x}; M_m)\big)$ be the target confidence when only the frames selected by M_m are kept, and $p_r = p_{v^*}\big(P(\mathbf{x}; \mathbf{1} - M_m)\big)$ the confidence when the selected frames are frozen and the complement is kept. Raw sufficiency $\mathrm{Suf}_{\mathrm{raw}} = p_k/p_o$ measures how much of the original confidence is preserved, but does not distinguish confidence attributable to motion from confidence already available in the static reference. For motion-identifiable clips we therefore define

$$\mathrm{Suf}_{\mathrm{hr}} = \frac{p_k - p_s}{h}, \qquad \mathrm{Nec}_{\mathrm{hr}} = \frac{p_o - p_r}{h} \tag{2}$$

$\mathrm{Suf}_{\mathrm{hr}}$ measures the fraction of motion-dependent confidence retained by the selected frames, and $\mathrm{Nec}_{\mathrm{hr}}$ the fraction removed when they are frozen. Both scores divide by h and become ill-conditioned as $h \to 0$. For static-dominant clips, no temporal subset can be reliably identified. FAM then reports the clip as static-dominant rather than forcing an arbitrary temporal explanation.

3.3 Appearance objective

Minimizing the negative log-probability of the target class (Li et al., 2021; Vasu et al., 2026) rewards masks that preserve the prediction, but does not test whether the selected content is more informative than its complement. Following iGOS++ (Khorram et al., 2021) and Xing et al. (2025), we train M_a with a two-sided probability gap on the noun head,

$$\mathcal{L}^{n}_{\text{pred}} = p_{n^*}\big(\hat{\mathbf{x}}^{\text{del}}\big) - p_{n^*}\big(\hat{\mathbf{x}}^{\text{ins}}\big). \tag{3}$$

Minimizing Eq. (3) maximizes the insertion score, so that the selected region alone recovers the prediction (*sufficiency*), and minimizes the deletion score, so that removing the region suppresses it (*necessity*). As a difference of probabilities in $[-1, 1]$, the objective is bounded and differentiable for every mask. The full appearance objective is

$$\mathcal{L}_{\text{app}} = \mathcal{L}^{n}_{\text{pred}} + \lambda_A \, \|M_a\|_1/|\Omega| + \lambda_B \, \mathcal{L}_{\text{BTV}}, \tag{4}$$

where $|\Omega| = THW$. The ℓ_1 term is a budget-free sparsity penalty, and $\mathcal{L}_{\text{BTV}}$ is an edge-aware bilateral total variation (Farsiu et al., 2004; Tomasi & Manduchi, 1998). We use $\mathcal{L}_{\text{BTV}}$ instead of the fixed-shape prior of Li et al. (2021), which encourages blob-like masks independently of image content. M_a is parameterized as a low-resolution seed that is upsampled and smoothed (Appendix A.1).

3.4 Motion objective

Headroom-constrained sufficiency. For motion-identifiable clips, we require the selected frames to recover at least a fraction τ_s of the headroom, $\mathrm{Suf}_{\mathrm{hr}} \geq \tau_s$. We impose this constraint through a smooth hinge,

$$\mathcal{L}^{v}_{\text{suf}} = \beta_s \, \mathrm{softplus}\left(\frac{\tau_s - \mathrm{Suf}_{\mathrm{hr}}}{\beta_s}\right), \tag{5}$$

which is close to zero once the constraint is satisfied and grows approximately linearly when it is violated. We use $\tau_s = 0.95$ and $\beta_s = 0.05$. For single-head models, M_a and M_m explain the same target, so we additionally steer M_m away from frames whose static appearance alone supports the prediction. We define the per-frame headroom $\phi_t = \max(0,\, p_o - p^t_s)$. Note that $h = \min_t \phi_t$ whenever $h \geq 0$. ϕ_t is zero where frame t alone recovers the target confidence, and we encourage M_m to align with it:

$$\mathcal{L}_{\text{static}} = 1 - \frac{M_m \cdot \phi}{\|M_m\| \, \|\phi\|}. \tag{6}$$

The full motion objective is

$$\mathcal{L}_{\text{mot}} = \lambda_s \mathcal{L}^{v}_{\text{suf}} + \lambda_M \|M_m\|_1/T + \lambda_{\text{stat}} \mathcal{L}_{\text{static}}. \tag{7}$$

For dual-head models, $\lambda_{\text{stat}} = 0$. The sparsity term removes frames while the sufficiency constraint remains satisfied, and removal stops when freezing further frames would bring $\mathrm{Suf}_{\mathrm{hr}}$ below τ_s. We do not optimize necessity. Because video frames are redundant, requiring high necessity forces M_m to cover every frame that could support the prediction, which in our preliminary experiments expanded M_m toward most of the clip. $\mathrm{Nec}_{\mathrm{hr}}$ is instead measured only after optimization, so that it remains an independent diagnostic of temporal reliance. We optimize M_m and M_a *independently* to

be budget-free and better aligned to object tracks as opposed to joint optimization, budget-dependant and video-independent structural prior as in (Vasu et al., 2026). M_a is not restricted to the frames selected by M_m, we compare with a restricted variant in Appendix A.8. Implementation details and hyperparameters are given in Appendix A.1.

4 Reliance Probes and Static-Bias Index

Action recognizers are known to exploit *static shortcuts*, recognizing a verb from the appearance of a single frame rather than from motion across time (Kowal et al., 2022; Li et al., 2023; Ilic et al., 2022). This behavior motivated temporally demanding benchmarks such as Something-Something (Goyal et al., 2017), in which distinguishing between action classes requires motion as evidence. The static reference and blur perturbation from §3 give a direct way to measure how much a trained model relies on static appearance for each prediction. We use them to define reliance probes for individual clips and a Static-Bias Index that summarizes these probes across clips, and use both to compare static and temporal reliance across architectures and datasets.

Reliance probes. For each clip, we use the prediction on the original clip as the target and measure how its confidence changes when temporal variation or appearance is removed. Temporal variation is removed with the static reference $\mathbf{x}^{\text{stat}}$, and appearance with the fully blurred clip $\tilde{\mathbf{x}}$. For target $c \in \{v^*, n^*\}$,

$$R^c_{\text{mot}} = \frac{p_c(\mathbf{x}) - p_c(\mathbf{x}^{\text{stat}})}{p_c(\mathbf{x})}, \qquad R^c_{\text{app}} = \frac{p_c(\mathbf{x}) - p_c(\tilde{\mathbf{x}})}{p_c(\mathbf{x})}, \tag{8}$$

with both quantities clipped to $[0, 1]$. $R^{v^*}_{\text{mot}}$ and $R^{n^*}_{\text{app}}$ measure the expected motion–verb and appearance–noun relationships. $R^{v^*}_{\text{app}}$ and $R^{n^*}_{\text{mot}}$ measure the cross-axis relationships, since a model can also use appearance to support its verb prediction or motion to support its noun prediction. A large $R^{v^*}_{\text{mot}}$ indicates that the verb prediction depends strongly on motion. For single-head models, $v^* = n^*$ and the probes reduce to R_{mot} and R_{app}.

Static-Bias Index. To summarize static reliance across a collection of clips, we define the SBI as

$$\text{SBI} = 1 - \mathbb{E}_{\text{clips}}\left[R^{v^*}_{\text{mot}}\right] \in [0, 1]. \tag{9}$$

A high SBI indicates that removing temporal variation causes little reduction in confidence, while a low SBI indicates that predictions depend strongly on temporal variation. Note that $R^{v^*}_{\text{mot}} = h/p_o$ before clipping, so the SBI is one minus the mean normalized motion headroom. The same static reference therefore determines both when a temporal explanation is identifiable (§3.2) and how strongly a model relies on static appearance across clips. We also report the *low-headroom rate*, the fraction of clips with $h < \tau_h$, which counts the predictions for which no temporal explanation can be identified.

Explanation-conditioned SBI. Eq. (8) removes temporal variation from the entire clip. It measures whether the model depends on motion overall, but cannot show whether this dependence is concentrated in a few frames or distributed redundantly across several subsets of frames. We test this by freezing only the frames that an explanation $\mathcal{E}$ selects, keeping the complementary frames unchanged, and measuring the drop in verb confidence,

$$\text{Nec}^v_{\text{ret}} = \frac{p_o - p_r}{p_o}, \tag{10}$$

where $p_r = p_{v^*}\big(P(\mathbf{x}; \mathbf{1} - M_m)\big)$ as in §3.2. Unlike Nec_{hr}, this drop is normalized by the clean confidence p_o rather than by the headroom, so it remains defined on static-dominant clips and these clips are not removed from the average. The explanation-conditioned SBI is

$$\text{SBI}_{\mathcal{E}} = 1 - \mathbb{E}_{\text{clips}}\left[\text{Nec}^v_{\text{ret}}\right]. \tag{11}$$

The SBI freezes the entire clip and measures the model's total reliance on temporal variation, whereas $\text{SBI}_{\mathcal{E}}$ freezes only the frames selected by an explanation method,$\mathcal{E}$ and measures how much of this reliance those frames account for. When $\mathcal{E}$ selects every frame, $P(\mathbf{x}; \mathbf{0}) = \mathbf{x}^{\text{stat}}$ and the two indices coincide. If $\text{SBI}_{\mathcal{E}}$ is close to SBI, the selected frames account for most of the model's temporal

reliance. If $\mathrm{SBI}_{\mathcal{E}}$ is higher than SBI, the model relies on motion, but freezing only the selected frames removes little of that reliance, because motion in other frames can support the same prediction. In this case, a clip may have more than one sufficient temporal explanation (Shitole et al., 2021), which we discuss in the limitations. $\mathrm{SBI}_{\mathcal{E}}$ can also fall below SBI because freezing selected frames may remove frame-specific appearance in addition to temporal variation. Such a gap should therefore not be interpreted as additional temporal reliance. Figure 1 demonstrates how headroom, reliance and SBI can be used.

5 Evaluation and Results

We evaluate static and temporal reliance on 200 clips each from EGTEA Gaze+ (Li et al., 2018a) and Something-Something V2 (Goyal et al., 2017). EGTEA is an egocentric dataset with synchronized gaze and verb–noun labels, while SSv2 contains 174 single-label classes designed around temporal changes. To examine whether reliance varies across architectures, we evaluate TSM (Lin et al., 2019) with a ResNet-50 backbone, TimeSformer (Bertasius et al., 2021), and SlowFast (Feichtenhofer et al., 2019). On EGTEA, all three use separate verb and noun heads. A single-frame ResNet with no temporal modeling provides a static control. For each model, we report the appearance and motion Reliance Probes, SBI, explanation-conditioned SBI, and low-headroom rate.

We then evaluate whether FAM's factored explanations recover the evidence identified by this audit. We compare against Extremal Perturbation (EP) (Fong et al., 2019) and STEP (Li et al., 2021). Since both produce a single spatiotemporal mask, we obtain EP-TD and STEP-TD by averaging their masks within each frame and thresholding at 0.5. We also include a *Unified Loss* that retains separate masks but optimizes both for the averaged noun–verb target while On single-head SSv2 this target averaging is identical to FAM, so we instead report *Unified Mask*, which uses a single conflated mask. Where feasible, an exact motion oracle evaluates all 2^T frame subsets and returns the smallest subset satisfying $\mathrm{Suf}_{\mathrm{hr}} \geq \tau_s$ while it is infeasible for SlowFast ($T = 40$).

Data	Model	R^v_{mot}	R^v_{app}	R^n_{mot}	R^n_{app}	SBI	$\mathrm{SBI}_{\mathrm{FAM}}$	$\mathrm{SBI}_{\mathrm{STEP}}$	$\mathrm{SBI}_{\mathrm{EP}}$	Low-hr%
EGTEA	RSingle	0	0.69	0.21	0.87	1.0	0.94	0.82	0.83	100
	TSM	0.13	0.75	0.47	0.93	0.87	0.81	0.67	0.58	70.5
	TSformer	0.15	0.72	0.50	0.89	0.85	0.84	0.75	0.83	68.5
	SlowFast	0.54	0.78	0.67	0.91	0.46	0.71	0.63	0.80	21.0
SSv2	RSingle	0.02	0.90	–	–	0.98	0.86	0.71	0.84	97.5
	TSM	0.87	0.93	–	–	0.13	0.33	0.50	0.42	8.0
	TSformer	0.86	0.89	–	–	0.14	0.51	0.55	0.56	8.5
	SlowFast	0.88	0.79	–	–	0.12	0.73	0.66	0.83	8.0

Table 1: **Appearance and motion reliance across datasets and architectures.** R_{mot} and R_{app} (Eq. 8) measure confidence lost when temporal variation or appearance is removed; v/n denote verb/noun heads, while SSv2 is single-head. SBI (Eq. 9) measures whole-clip static reliance, with higher values indicating greater static reliance. $\mathrm{SBI}_{\mathrm{FAM}}$, $\mathrm{SBI}_{\mathrm{STEP}}$, and $\mathrm{SBI}_{\mathrm{EP}}$ (Eq. 11) freeze only explainer-selected frames; values closer to SBI indicate that the selected frames capture more of the model's temporal reliance. Low-hr is the percentage of clips with $h < \tau_h$, where temporal attribution is unidentifiable. RSingle is a single-frame static control.

Our evaluation asks three questions: (1) how static and temporal reliance vary across datasets and architectures; (2) how much of this temporal reliance is captured by explanation-selected frames; and (3) when temporal attribution is identifiable, how well FAM separates and localizes the evidence associated with each axis. Because standard insertion/deletion metrics (Petsiuk et al., 2018) score a single mask against one class probability, we evaluate the two FAM axes separately. Appearance is scored by insertion and deletion AUC on the noun (or action class for single-head models) over all $T \times H \times W$ locations relative to the blurred clip $\tilde{\mathbf{x}}$. Motion is scored on the verb using $\mathrm{Suf}_{\mathrm{hr}}$ and $\mathrm{Nec}_{\mathrm{hr}}$ (Eq. (2)) under the replacement operator P, restricted to motion-identifiable clips; static-dominant clips are instead reported through the low-headroom rate. We also report the mean active-frame fraction $\overline{M}_m$. All temporal baselines are evaluated with the same operator P. Implementation details are given in Appendix A.1.

5.1 Static-bias audit

Table 1 reports the reliance probes, SBI, and low-headroom rate for each model and dataset. On SSv2, TSM, TimeSformer, and SlowFast have similar motion reliance ($R^v_{\mathrm{mot}} = 0.86$–$0.88$), SBI ($0.12$–$0.14$), and low-headroom rates (8.0–8.5%). On EGTEA, the corresponding values vary more widely: $R^v_{\mathrm{mot}} = 0.13$–$0.54$, SBI 0.46–0.87, and low-headroom 21.0–70.5%. This difference is consistent with the datasets themselves. SSv2 classes are defined largely by transformations and

temporal relations, whereas EGTEA actions often contain object and scene cues that can support verb prediction from individual frames. Architecture has a larger effect on EGTEA: SlowFast reaches $R^v_{\text{mot}} = 0.54$, while TSM and TimeSformer remain near 0.14. SlowFast separates low-rate spatial processing from a high-rate pathway designed to capture rapidly changing motion, whereas TSM mixes neighboring-frame features through channel shifts and TimeSformer learns temporal interactions through attention. When temporal information is required, as on SSv2, all three models use it. When static cues provide an alternative, as on EGTEA, SlowFast's high-rate pathway is associated with greater use of temporal evidence. We treat this as an empirical association rather than evidence that the architecture alone causes the difference. The single-frame control behaves as expected, with SBI of $1.00/0.98$ and low-headroom rates of $100/97.5\%$ on EGTEA/SSv2, providing a reference for predictions supported almost entirely by static information. These findings remain stable under calibration-insensitive SBI variants and τ_h (Appendices A.4 and A.5).

Appearance remains important when motion reliance is high. R_{app} ranges from 0.69 to 0.93 across all models, so the more informative pattern is the relationship between appearance and motion reliance. For example, EGTEA–TSM has $R^v_{\text{app}} = 0.75$ and $R^v_{\text{mot}} = 0.13$, whereas SSv2–TSM has 0.93 and 0.87, showing much greater reliance on motion on SSv2. On EGTEA, noun predictions rely most strongly on appearance ($R^n_{\text{app}} = 0.87$–0.93), and are also reduced when temporal variation is removed ($R^n_{\text{mot}} = 0.47$–0.67 for the temporal models). This noun probe should be read with one caveat. $\mathbf{x}^{\text{stat}}$ repeats the frame that maximizes the verb, not the noun, so part of the drop in noun confidence can come from which frame is repeated rather than from the removed motion. The single-frame control shows this effect: it has no temporal modeling, yet $R^n_{\text{mot}} = 0.21$ on EGTEA. We therefore treat R^n_{mot} as an upper estimate of the noun's reliance on motion. These cross-axis effects support conditioning each mask on its own prediction head while evaluating both types of evidence against both heads.

Explanation-conditioned SBI. We compare each explainer's $\text{SBI}_{\mathcal{E}}$ with the whole-clip SBI to assess whether the selected frames account for the model's temporal reliance. On the temporal models other than SlowFast, FAM gives the closest agreement in all four settings, with a mean absolute gap of 0.16, compared with 0.26 for EP and 0.27 for STEP. The two exceptions are SlowFast, where STEP is closer. One possible explanation is that the slow pathway is formed from a temporally subsampled version of the input, so freezing one temporal position may leave related information available elsewhere in the representation. On SSv2–SlowFast, for example, SBI is 0.12 while SBI_{FAM} is 0.73. In such cases, the whole-clip probe provides the more direct measure of temporal reliance, and comparisons between explainers remain meaningful within the same model–dataset setting.

5.2 Explanation quality

Table 2 evaluates FAM's appearance and motion explanations under the separation protocol, over three runs. FAM achieves the highest insertion and lowest deletion AUC in all six settings and improves over Unified on EGTEA. For example, SlowFast insertion increases from 0.64 to 0.69 while deletion remains 0.10. Appearance results are stable across runs (insertion SD ≤ 0.02, deletion SD ≤ 0.01); paired TSM tests are reported in Appendix 11. These results indicate that averaging the noun and verb targets weakens the appearance explanation, and that optimizing M_a for the noun lets it focus on appearance evidence.

FAM finds temporally sufficient frame sets. Across all six settings, FAM achieves $\text{Suf}_{\text{hr}} = 0.93$–$0.99$, compared with 0.50–0.74 for the best temporal baseline on EGTEA and 0.48–0.70 on SSv2, with SD below 0.02. FAM has the highest practical necessity in five of six settings; EP-TD is higher on EGTEA–TSM (0.78 versus 0.70). Lower necessity than sufficiency is consistent with redundant temporal evidence elsewhere in a clip.

High sufficiency does not imply the smallest temporal mask. FAM generally uses more frames than the temporal baselines, although those smaller baseline masks also have much lower sufficiency. The exact oracle shows that smaller masks can reach similar sufficiency: in the four settings where it is available, the oracle retains 28–40% of the frames, whereas FAM retains 52–66%. FAM therefore reliably satisfies its sufficiency constraint, but its gradient-based optimization does not always find the smallest mask that satisfies it. The standard deviations show the same distinction: sufficiency

(a) Appearance explanations					(b) Motion explanations			
Data	**Model**	**Method**	**Ins↑**	**Del↓**	**Method**	**Suf$_{hr}$ ↑**	**Nec$_{hr}$ ↑**	**M_m% ↓**
EGTEA	TSM	Extremal	0.588 ± 0.008	0.236 ± 0.002	Oracle	0.983 ± 0.001	0.563 ± 0.033	33.55 ± 1.65%
		STEP	0.665 ± 0.003	0.299 ± 0.019	Ext.-Temp	0.738 ± 0.050	**0.776** ± 0.041	53.63 ± 0.22%
		Unified Loss	0.799 ± 0.010	0.119 ± 0.001	STEP-TD	0.564 ± 0.071	0.645 ± 0.051	**45.96** ± 1.63%
		FAM	**0.823** ± 0.002	**0.114** ± 0.002	FAM	**0.940** ± 0.018	0.700 ± 0.049	56.04 ± 3.83%
	TSF	Extremal	0.680 ± 0.011	0.277 ± 0.024	Oracle	0.991 ± 0.004	0.440 ± 0.080	27.78 ± 1.98%
		STEP	0.718 ± 0.012	0.410 ± 0.028	Ext.-Temp	0.413 ± 0.080	0.476 ± 0.096	**44.55** ± 1.15%
		Unified Loss	0.813 ± 0.009	0.134 ± 0.001	STEP-TD	0.500 ± 0.035	0.514 ± 0.088	47.08 ± 2.05%
		FAM	**0.839** ± 0.007	**0.126** ± 0.009	FAM	**0.931** ± 0.002	**0.665** ± 0.048	52.27 ± 6.19%
	SF	Extremal	0.577 ± 0.010	0.166 ± 0.010	Oracle	–	–	–
		STEP	0.596 ± 0.010	0.259 ± 0.010	Ext.-Temp	0.390 ± 0.036	0.247 ± 0.019	**33.55** ± 0.50%
		Unified Loss	0.643 ± 0.028	0.102 ± 0.002	STEP-TD	0.699 ± 0.046	0.527 ± 0.029	49.05 ± 0.71%
		FAM	**0.686** ± 0.016	**0.100** ± 0.002	FAM	**0.937** ± 0.014	**0.575** ± 0.022	53.44 ± 2.17%
SSv2	TSM	Extremal	0.300 ± 0.022	0.122 ± 0.012	Oracle	0.988 ± 0.003	0.550 ± 0.012	39.92 ± 0.41%
		STEP	0.325 ± 0.022	0.138 ± 0.012	Ext.-Temp	0.464 ± 0.054	0.592 ± 0.012	**46.79** ± 0.13%
		Unified Mask	0.306 ± 0.023	0.257 ± 0.017	STEP-TD	0.484 ± 0.021	0.545 ± 0.026	50.13 ± 0.61%
		FAM	**0.496** ± 0.021	**0.071** ± 0.003	FAM	**0.989** ± 0.002	**0.748** ± 0.028	65.77 ± 0.32%
	TSF	Extremal	0.471 ± 0.015	0.142 ± 0.009	Oracle	0.988 ± 0.002	0.312 ± 0.025	31.39 ± 0.71%
		STEP	0.512 ± 0.010	0.202 ± 0.012	Ext.-Temp	0.615 ± 0.051	0.491 ± 0.005	**48.71** ± 0.73%
		Unified Mask	0.436 ± 0.024	0.355 ± 0.012	STEP-TD	0.623 ± 0.022	0.495 ± 0.036	51.17 ± 0.90%
		FAM	**0.642** ± 0.010	**0.094** ± 0.007	FAM	**0.986** ± 0.002	**0.578** ± 0.033	62.44 ± 2.94%
	SF	Extremal	0.253 ± 0.017	0.076 ± 0.009	Oracle	–	–	–
		STEP	0.240 ± 0.014	0.105 ± 0.008	Ext.-Temp	0.416 ± 0.035	0.154 ± 0.013	**30.96** ± 0.62%
		Unified Mask	0.263 ± 0.008	0.273 ± 0.015	STEP-TD	0.704 ± 0.035	0.335 ± 0.030	50.31 ± 0.42%
		FAM	**0.395** ± 0.014	**0.068** ± 0.010	FAM	**0.958** ± 0.011	**0.455** ± 0.030	46.56 ± 1.89%

Table 2: **Comparison of appearance and motion explanations across datasets and recognition architectures.** Values are mean ± standard deviation across 3 trials on 200 clips per dataset. TSF denotes TimeSformer and SF denotes SlowFast. Motion metrics are computed on motion-identifiable clips only ($h \geq \tau_h$); the fraction of such clips for each model is given by the low-headroom rate in Table 1. The best practical method within each dataset/model setting is shown in **bold**. The exact motion oracle returns the smallest frame subset with $\mathrm{Suf}_{\mathrm{hr}} \geq \tau_s$ and is excluded from bolding because it is a reference rather than a practical explanation method. It is infeasible for SlowFast ($T = 40$). Unified loss is reported only for EGTEA; with SSv2's single prediction head, averaging the targets gives the FAM objective while Unified Mask is only reported for SSV2.

varies little across runs, whereas $\overline{M}_m$ varies by as much as 6.19 percentage points. Different runs can therefore reach similarly sufficient solutions with different numbers of frames. To check if high sufficiency is simply the result of larger masks, we include a size-matched random control in Appendix A.6, showing that the selected temporal locations outperform random subsets of the same size.. Qualitative examples of both masks are shown in Appendix A.2.

6 Conclusion

We studied how video action recognition models rely on static appearance versus temporal variation. Motion headroom determines when temporal attribution is identifiable, while Reliance Probes and the Static-Bias Index quantify static reliance across predictions and models. The explanation-conditioned SBI further measures how much of this reliance is captured by an explanation's selected frames. Across EGTEA Gaze+ and Something-Something V2, TSM, TimeSformer, and SlowFast show greater static reliance on EGTEA and stronger temporal reliance on SSv2, a pattern that remains stable under calibration-insensitive measures and across headroom thresholds. Together, these tools provide instance-level appearance and temporal explanations and a post-hoc audit of static reliance.

Limitations and future work. The Reliance Probes and SBI are post-hoc, perturbation-based diagnostics rather than causal measures: their values depend on how appearance and temporal variation are removed and on the static reference used to define temporal reliance. The cross-axis probes further show that appearance–noun and motion–verb dependence is not exclusive. At the explanation level, FAM inherits this perturbation dependence, identifies *when* but not where or in which direction motion matters, can return larger-than-minimal sufficient masks, and produces only one explanation even when alternatives may exist (Shitole et al., 2021). Future work can study class-level reliance, additional datasets, architectures and modalities, alternative interventions for measuring static and temporal reliance, and richer temporal explanations that localize motion and recover multiple minimal solutions.

REFERENCES

Gedas Bertasius et al. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

Jinwoo Choi et al. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in neural information processing systems*, 32, 2019.

D. Damen et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Jay DeYoung et al. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4443–4458, 2020.

Sina Farsiu et al. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.

C. Feichtenhofer et al. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

R. Fong et al. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Ruth C Fong et al. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 3449–3457. IEEE, 2017.

R. Goyal et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

T. Hartley et al. Swag-v: Explanations for video using superpixels weighted by average gradients. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

De-An Huang et al. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 7366–7375. IEEE, 2018.

Filip Ilic et al. Is appearance free action recognition possible? In *European Conference on Computer Vision*, pp. 156–173. Springer, 2022.

Y. Ji et al. Spatial-temporal concept based explanation of 3d convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Saeed Khorram et al. iGOS++: Integrated gradient optimized saliency by bilateral perturbations. In *Proceedings of the Conference on Health, Inference, and Learning*, 2021.

M. Kowal et al. Understanding video transformers via universal concept discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Matthew Kowal et al. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 13999–14009, 2022.

J. Lee et al. Disentangled concepts speak louder than words: Explainable video action recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Haoxin Li et al. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19911–19923, 2023.

Y. Li et al. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018a.

Yingwei Li et al. Resound: Towards action recognition without representation bias. In *European conference on computer vision*, pp. 520–535. Springer, 2018b.

Z. Li et al. Towards visually explaining video understanding networks with perturbation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

J. Lin et al. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Z. Liu et al. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

V. Petsiuk et al. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

A. Saha et al. Exploring explainability in video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

R. R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Laura Sevilla-Lara et al. Only time can tell: Discovering temporal data for temporal modeling. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 535–544. IEEE, 2021.

Vivswan Shitole et al. One explanation is not enough: structured attention graphs for image classification. *Advances in Neural Information Processing Systems*, 34:11352–11363, 2021.

A. Stergiou et al. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

Pascal Sturmfels et al. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

Carlo Tomasi et al. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pp. 839–846. IEEE, 1998.

Natalia Valderrama et al. Towards holistic surgical scene understanding. In *International conference on medical image computing and computer-assisted intervention*, pp. 442–452. Springer, 2022.

Bhavan Vasu et al. Disentangling where and when: Factored spatio-temporal explanations for video action recognition. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026.

H. Wang et al. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

Xiaoying Xing et al. Where do large vision-language models look at when answering questions? *arXiv preprint arXiv:2503.13891*, 2025.

T. Zhuo et al. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the ACM International Conference on Multimedia*, 2019.

A Appendix

A.1 Implementation details.

Each stage (M_m and M_a) runs 500 Adam steps at learning rate 0.05 on $T = 8$ uniformly sampled frames at 224×224 for TSM and TimesFormer and $T = (32 + 8) = 40$ for SlowFast Network; Motion mask stage additionally decays the learning rate linearly to zero over its final 20%. The four loss weights $(\lambda_A, \lambda_M, \lambda_{BTV}, \lambda_{stat})$ are set using the composite objectives in section A.7. Full set of hyperparameter values are presented in Appendix Table 12. Code and trained checkpoints will be released upon acceptance.

A.2 Qualitative Results

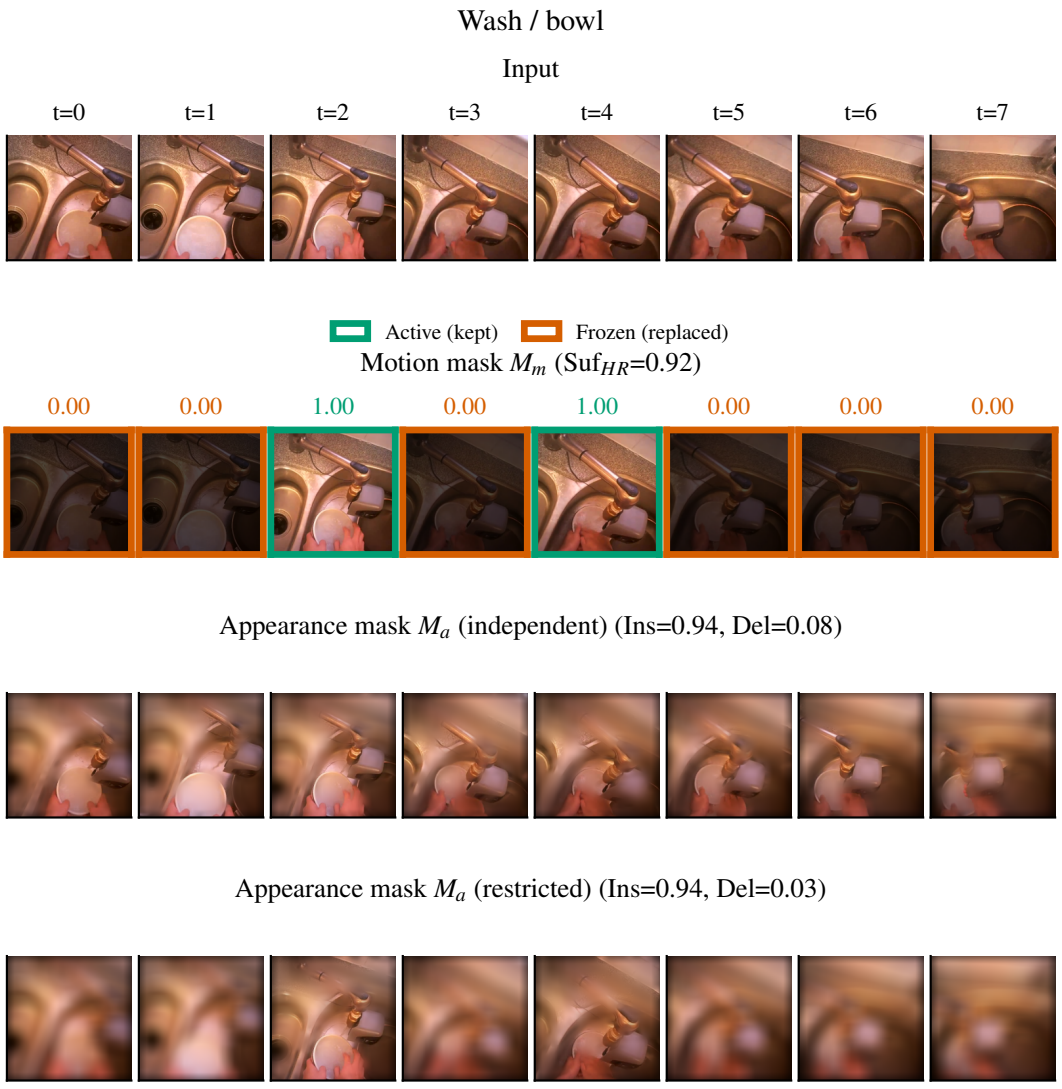


Figure 2: Qualitative example for class "Wash Bowl" from the EGTEA dataset.

Figure 2 and 3 show qualitative figures for examples from the EGTEA and SSv2 dataset respectively. The first row is the original video frames, followed by the motion mask M_m with active frames marked in green. The third row the appearance mask M_a followed by the motion restricted appearance mask in the last row. Figure 4 shows how SBI and Reliance is spread for two examples from the EGTEA and SSV2 dataset using a TSM model.

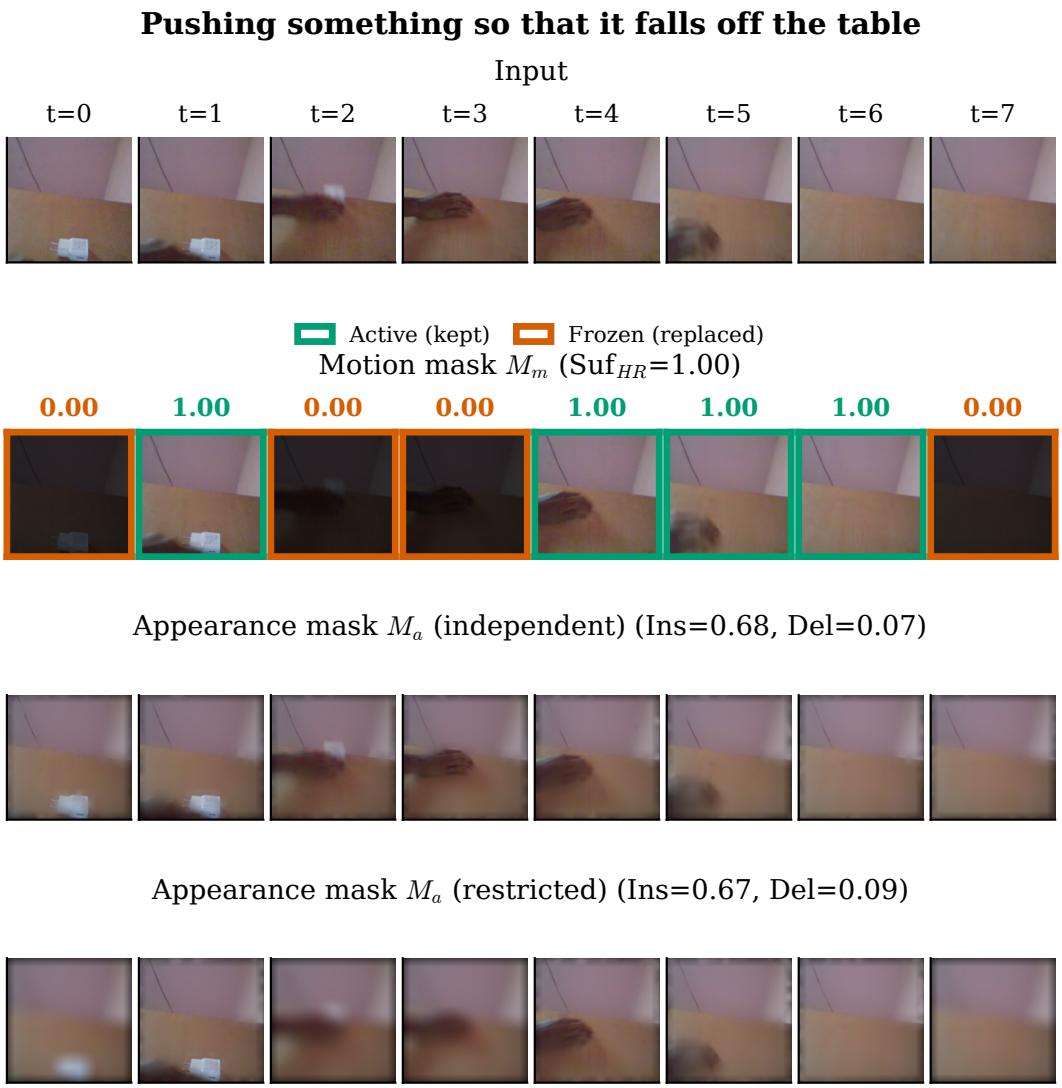


Figure 3: Qualitative example for class “Push something so that it falls off the table" from SSv2 dataset.

A.3 Temporal factor: nearest-frame replacement.

For completeness, we describe the nearest-frame replacement perturbation strategy(Vasu et al., 2026) for perturbing motion from a video while preserving the visual content of its frames. The strategy is controlled by a temporal mask $M_m = \{m_t\}_{t=1}^{T}$, where T is the number of timesteps in the video and each value m_t indicates whether the corresponding frame remains active. Frames marked inactive are replaced with unaltered copies of nearby active frames from the original clip x. Repeating an active frame at neighboring time steps locally pauses the video, thereby supressing temporal changes without introducing out-of-distribution colors, textures or objects.

Implementing this perturbation requires resolving two design hurdles. First, the forward computation must make a binary decision at each time step: a frame must either remain unchanged ($m_t = 1$) or be replaced by the content of another frame ($m_t = 0$). Second, the temporal mask must support gradient-based optimization so that the importance assigned to each frame can be learned from the model's response. We therefore use a discrete temporal mask during the forward computation and a continuous version of the same mask during gradient computation using Straight-Through Estimator given by:

$$M_{m_t}^{\text{STE}} = \mathbb{1}[m_t > 0.5] - \text{sg}(m_t) + m_t, \tag{12}$$

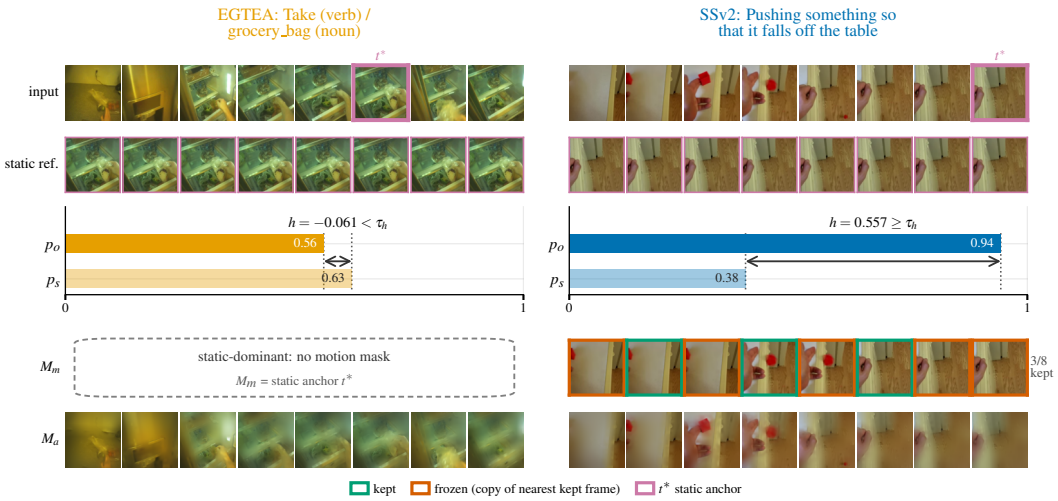


Figure 4: Qualitative example showing the use of headroom along with its motion and appearance explanations from FAM.

where $\mathbb{1}[\cdot]$ is the indicator function and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation.

Once a frame is marked inactive, the perturbation operator must determine which active frame should replace it. Candidate replacements j are assigned larger weights $\tilde{w}_{j\to i}$ when they are temporally closer to the inactive frame i, while frames marked inactive are excluded from consideration as a potential replacement. This modeled preference for nearby active frames limits large changes in visual content and produces local pauses rather than abrupt transitions between active and inactive frames.

The distance based replacement weights are computed from:

$$w_{j\to i} = \frac{e^{d(i,j)} \cdot {M_m}_j^{\mathrm{STE}}}{\sum_{k\neq i} e^{d(i,k)} \cdot {M_m}_k^{\mathrm{STE}} + \epsilon}. \tag{13}$$

The exponential term assigns larger weights to frames that are temporally closer to i using $d(i,j) = -|i-j|^2/2\sigma_R^2$, while progressively reducing the contribution of more distant frames. $\sigma_R = 2$ controls the rate at which the preference for replacement frames decreases with temporal distance between a replacement candidate j and the inactive frame i. Multiplication by ${M_m}_j^{\mathrm{STE}}$ restricts the candidate set to active frames: when frame j is inactive, ${M_m}_j^{\mathrm{STE}} = 0$, and its replacement weight becomes zero. Thus, $w_{j\to i}$ represents a continuous preference over active replacement frames and, selecting a weighted average of these candidates would combine content from several time steps and could introduce ghosting artifacts. Therefore, the candidate with the largest weight in the forward computation must be selected and a second straight-through estimator is used to propagate gradients through the continuous weights $w_{j\to i}$ as shown below.

$$j^*(i) = \arg\max_j w_{j\to i},$$

$$\tilde{w}_{j\to i} = \mathbb{1}[j = j^*(i)] - \mathrm{sg}(w_{j\to i}) + w_{j\to i},$$

The final perturbed clip is

$$\hat{\mathbf{x}}_t = {M_m}_t^{\mathrm{STE}} \cdot \mathbf{x}_t + (1 - {M_m}_t^{\mathrm{STE}}) \cdot \sum_j \tilde{w}_{j\to t}\, \mathbf{x}_j. \tag{14}$$

where the first term retains the original frame when ${M_m}_t^{\mathrm{STE}} = 1$, and the second term replaces it when ${M_m}_t^{\mathrm{STE}} = 0$. Because $\tilde{w}_{j\to i}$ is one-hot during the forward computation, the replacement is an exact copy of a single active frame. It is possible for the binary temporal mask ${M_m}^{\mathrm{STE}}$ to contain no active frames. In this case, the perturbation uses a static fallback constructed from the single frame that produces the highest confidence for the target verb when repeated throughout the video as shown below:

Data	Model	$\mathrm{SBI}_{\mathrm{prob}}$	$\mathrm{SBI}_{\mathrm{marg}}$	$\mathrm{SBI}_{\mathrm{flip}}$	$\rho_{\mathrm{p,m}}$	Low-hr%
EGTEA	RSingle	1.00	0.983	0.990	$+0.26$	100.0
	TSM	0.87	0.750	0.930	$+0.97$	70.5
	TimeSformer	0.85	0.731	0.860	$+0.94$	68.5
	SlowFast	0.46	0.276	0.420	$+0.85$	21.0
SSv2	RSingle	0.98	0.854	0.915	$+0.27$	97.5
	TSM	0.13	0.085	0.115	$+0.55$	8.0
	TimeSformer	0.14	0.128	0.200	$+0.69$	8.5
	SlowFast	0.12	0.076	0.145	$+0.61$	8.0

Table 3: **Calibration robustness of static-bias measurements.** Results are reported over 200 clips per row. $\rho_{\mathrm{p,m}}$ is the Spearman correlation between the per-clip probability- and margin-based motion-reliance scores. Because the three SBI variants use different quantities, values should be compared within each column rather than directly across columns.

$$a^* = \arg\max_{t\in\{1,\dots,T\}} p_{v^*}\big(\text{Repeat}(x_t, T)\big),$$
$$\mathbf{x}^{\text{stat}} = \text{Repeat}(x_{a^*}, T).$$

The resulting clip contains only frames drawn directly from the original video. Replaced frames therefore preserve the colors, textures, objects, and spatial arrangement present in the input, while perturbing motion dynamics.

A.4 Calibration robustness of the Static-Bias Index

The original motion-reliance probe R^v_{mot} measures the normalized drop in softmax probability after temporal variation is removed. Because softmax probabilities depend on model calibration, two models with similar dependence on motion could produce different SBI values if their confidence distributions have different sharpness. The fixed headroom threshold $\tau_h = 0.05$ can be affected in the same way. We therefore evaluate the same static intervention using three measures:

$$\mathrm{SBI}_{\mathrm{prob}} = 1 - \mathbb{E}[(p_0 - p_s)/p_0]\,, \tag{15}$$
$$\mathrm{SBI}_{\mathrm{marg}} = 1 - \mathbb{E}[(m_0 - m_s)/|m_0|]\,, \quad m = z_t - \max_{j\neq t} z_j, \tag{16}$$
$$\mathrm{SBI}_{\mathrm{flip}} = 1 - \Pr[\arg\max \text{ changes}]. \tag{17}$$

Here, $\mathrm{SBI}_{\mathrm{prob}}$ is the probability-based measure used in the main paper, $\mathrm{SBI}_{\mathrm{marg}}$ uses the target-class logit margin, and $\mathrm{SBI}_{\mathrm{flip}}$ measures whether removing temporal variation changes the predicted class. Under temperature scaling $z \to z/T$, both margins are scaled by the same factor, leaving $\mathrm{SBI}_{\mathrm{marg}}$ unchanged. $\mathrm{SBI}_{\mathrm{flip}}$ is also unchanged by any recalibration that preserves the predicted class and does not require a confidence threshold.

The main conclusions from the probability-based SBI remain unchanged under the calibration-insensitive alternatives. First, EGTEA remains more static-dependent than SSv2. Among the temporal models, the lowest $\mathrm{SBI}_{\mathrm{prob}}$ on EGTEA is SlowFast at 0.46, which is still above the highest value on SSv2, TimeSformer at 0.14. The same separation is present for $\mathrm{SBI}_{\mathrm{marg}}$ (0.276 versus 0.128) and $\mathrm{SBI}_{\mathrm{flip}}$ (0.420 versus 0.200). Thus, the difference in static reliance between the two datasets is not explained by softmax calibration alone.

The architecture pattern on EGTEA is also preserved. Under all three measures, SlowFast shows greater temporal reliance than TSM and TimeSformer, while the single-frame model remains strongly static-dependent. Architecture also produces a larger spread on EGTEA than on SSv2. Across the three temporal models, $\mathrm{SBI}_{\mathrm{prob}}$ ranges by 0.41 on EGTEA but only 0.02 on SSv2. The same pattern appears for $\mathrm{SBI}_{\mathrm{flip}}$, with ranges of 0.51 and 0.085, respectively. The ordering of the closely grouped SSv2 models varies slightly across the measures, but all three remain substantially more motion-dependent than their EGTEA counterparts.

The absolute SBI values differ because the measures describe the same intervention using different quantities. $\mathrm{SBI}_{\mathrm{marg}}$ is generally lower than $\mathrm{SBI}_{\mathrm{prob}}$ because softmax can compress changes in logit margin when the original prediction is already highly confident. Nevertheless, the probability- and margin-based reliance scores are strongly correlated for the EGTEA temporal models, with $\rho_{\mathrm{p,m}} = 0.85$–$0.97$. The correlations are lower for the single-frame controls because their motion

Data	Model	Low-headroom % at τ_h				
		0.01	0.02	0.05	0.10	0.20
EGTEA	Single-frame	99.5	99.5	99.5	100.0	100.0
	TSM	66.5	67.0	70.5	73.0	79.5
	TimeSformer	61.5	63.0	68.5	74.5	80.0
	SlowFast	17.0	18.0	21.0	24.0	31.5
SSv2	Single-frame	95.5	96.5	98.5	99.5	100.0
	TSM	7.0	8.0	8.0	8.0	15.5
	TimeSformer	7.5	7.5	8.5	11.0	14.0
	SlowFast	6.0	7.5	8.0	9.0	12.0
$\min(\text{EGTEA}) - \max(\text{SSv2})$		+9.5	+10.0	+12.5	+13.0	+16.0

Table 4: **Sensitivity of the low-headroom rate to τ_h.** Results are reported over 200 clips per row. The final row compares the lowest low-headroom rate among the EGTEA temporal models with the highest rate among the SSv2 temporal models. Static single-frame controls are excluded from this comparison because they remain near 100% on both datasets.

reliance is close to zero for most clips, leaving little variation across examples. Across the six temporal models, SBI_{prob} and SBI_{flip} differ by only 0.035 on average. Together, these results show that the main static-bias findings are preserved when reliance is measured without depending on the calibration of softmax probabilities.

A.5 Sensitivity to the headroom threshold

The low-headroom rate is the only reported static-bias measure that depends on the threshold τ_h. SBI and its calibration-insensitive variants (§A.4) include all clips and do not use this threshold. Since the low-headroom rate measures the fraction of clips with $h < \tau_h$, increasing τ_h will naturally increase the reported rate. We therefore test whether the conclusions drawn from this measure remain stable as the threshold changes.

The difference between EGTEA and SSv2 remains at every threshold. Even the lowest low-headroom rate among the EGTEA temporal models is higher than the highest rate among the SSv2 models. The gap ranges from 9.5 to 16.0 percentage points as τ_h varies from 0.01 to 0.20. Thus, changing the threshold over a $20\times$ range changes the absolute rates but does not change the conclusion that static appearance supports substantially more predictions on EGTEA than on SSv2.

The architecture pattern on EGTEA is also stable. SlowFast has a substantially lower low-headroom rate than TSM and TimeSformer at every threshold, with a gap of at least 37 percentage points. TSM and TimeSformer remain much closer to one another and exchange order at some thresholds, so we do not interpret their small difference as meaningful. On SSv2, all three temporal architectures remain close, differing by at most a few percentage points across most of the sweep. Overall, the main dataset and architecture findings are therefore not sensitive to the particular choice of $\tau_h = 0.05$.

A.6 Random-mask control.

Table 5 tests whether FAM's motion results come from the selected frame locations rather than from mask size. Each FAM mask is compared with random masks containing the same number of active frames. FAM achieves higher raw sufficiency, headroom-normalized sufficiency, and necessity in all six settings. The gains are largest on average for headroom-normalized sufficiency ($+0.22$ to $+0.40$), while necessity gains are smaller but consistently positive ($+0.06$ to $+0.16$). Since mask size is fixed, these improvements show that FAM selects more informative temporal locations than random subsets of the same size. The smaller necessity gains are again consistent with temporal evidence being available in several parts of a clip. A matched-budget comparison with the temporal baselines is given in Appendix A.11.

A.7 Axis-specific loss-weight selection.

We search over the complete hyperparameter tuple $\boldsymbol{\lambda} = (\lambda_A, \lambda_M, \lambda_{BTV}, \lambda_{stat})$ by executing the procedure for every candidate configuration on a held-out validation subset. The spatial quality of

Dataset	Model	Suf_{raw} FAM	Suf_{raw} Rand.	Suf_{raw} $\Delta \uparrow$	Suf_{hr} FAM	Suf_{hr} Rand.	Suf_{hr} $\Delta \uparrow$	Nec FAM	Nec Rand.	Nec $\Delta \uparrow$
EGTEA	TSM	0.993	0.728	0.265	0.948	0.607	0.341	0.709	0.592	0.117
	TimeSformer	0.990	0.676	0.314	0.932	0.533	0.399	0.665	0.601	0.064
	SlowFast	0.974	0.745	0.229	0.954	0.739	0.215	0.583	0.497	0.086
SSv2	TSM	0.989	0.757	0.232	0.997	0.740	0.257	0.767	0.682	0.085
	TimeSformer	0.991	0.777	0.214	0.990	0.751	0.239	0.555	0.498	0.057
	SlowFast	0.960	0.713	0.247	0.959	0.694	0.265	0.485	0.325	0.160

Table 5: **Size-matched random-mask control.** FAM motion masks are evaluated against random masks of identical frame size with uniformly sampled locations (200 clips, 5 masks/clip). Δ denotes FAM minus random; positive values indicate that FAM-selected frames yield higher sufficiency or necessity than random subsets of the same size. Mask area is identical by construction.

each candidate hyperparameter tuple is evaluated using

$$\text{Comp}_a = \frac{1}{2}\left(\text{Ins}_a + (1 - \text{Del}_a)\right), \tag{18}$$

where higher insertion AUC indicates that the most highly attributed pixels rapidly recover the target prediction, and lower deletion AUC indicates that removing those pixels rapidly suppresses it. The two terms are weighted equally because they assess complementary properties of the same soft spatial ranking.

For the temporal axis, Suf_{hr} generally increases with mask size and can remain close to one i.e all active frames. Therefore maximizing it directly therefore favors unnecessarily large masks. Conversely, an objective of the form $\text{Suf}_{\text{hr}} - w\overline{M}_m$ introduces an arbitrary exchange rate between fidelity and parsimony. We instead use constrained minimality. Let

$$\mathcal{I} = \{i : p_o^{(i)} - p_s^{(i)} \geq \tau_h\} \tag{19}$$

denote the motion-identifiable validation clips, and define

$$S_i = \text{Suf}_{\text{hr}}^{(i)}, \qquad \overline{A_i} = \frac{1}{T}\sum_{t=1}^{T} m_t^{(i)}.$$

A candidate tuple is considered motion-feasible only if the mean, coverage, and lower tail of $S_{i\in\mathcal{I}}$ satisfy the prescribed validation thresholds τ_s. Among feasible candidates, we minimize the mean motion-mask area,

$$\overline{M}_m = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} A_i. \tag{20}$$

This criterion encodes the temporal explanation must first recover the required motion-dependent confidence, after which its support is made as small as possible. We perform grid search over $\boldsymbol{\lambda}$ to find the final hyperparameters.

A.8 Ablation: independent vs. motion-restricted appearance

By default, FAM optimizes the appearance mask M_a over all frames in the clip, independently of the motion mask M_m. We compare this design with a restricted variant in which appearance can only be selected from frames that are active in M_m:

$$M_a \leftarrow M_a \odot M_m, \qquad \text{applied every iteration, before any loss term.} \tag{21}$$

Under this restriction, M_a is exactly zero on frames that are inactive in M_m. Both variants use the same motion mask, so the comparison isolates the effect of restricting appearance to motion-selected frames. For the restricted variant, the appearance insertion and deletion evaluation also uses M_m to keep inactive frames frozen through $P(\cdot)$, matching the perturbation used during optimization.

Deletion changes little under the restriction. On EGTEA, deletion increases from 0.114 to 0.117, while on SSv2 it increases from 0.071 to 0.080. Thus, restricting the mask removes some selected appearance content without causing a large change in deletion performance.

Effect on insertion. The effect on insertion differs between the two datasets. On EGTEA, restricting M_a decreases insertion from 0.823 to 0.737, a difference of 0.086. On SSv2, insertion increases from 0.490 to 0.570, a difference of 0.080.

This difference is consistent with how the two datasets are modeled. SSv2 uses a single prediction head, so both M_m and M_a are optimized for the same target class. Restricting M_a to frames selected by M_m therefore keeps the appearance optimization within frames that were already selected for that prediction.

Table 6: Independent vs. motion-restricted appearance, TSM. Mean $\pm$ std over 3 seeds.

Dataset	Variant	App Ins↑	App Del↓
EGTEA	independent	**0.823** $\pm$ 0.002	**0.114** $\pm$ 0.002
	restricted	0.737 $\pm$ 0.005	0.117 $\pm$ 0.011
SSv2	independent	0.490 $\pm$ 0.021	**0.071** $\pm$ 0.003
	restricted	**0.570** $\pm$ 0.011	0.080 $\pm$ 0.003

EGTEA instead uses separate verb and noun heads. The motion mask M_m is optimized using the verb prediction, whereas the appearance mask M_a is optimized using the noun prediction. A frame that is not important for the verb can still contain useful information about the noun. Restricting M_a to the verb-selected frames can therefore remove appearance information that supports noun recognition, which is consistent with the lower insertion score on EGTEA.

These results support keeping M_a and M_m independent in the general formulation. Restricting appearance to motion-selected frames can be useful when both masks explain the same prediction, as on SSv2, but can remove useful appearance information when the masks explain different prediction heads.

Low-headroom clips. In this experiment, 70.5% of EGTEA clips and 8.0% of SSv2 clips have near-zero motion headroom. For a low-headroom clip, M_m reduces to the single static-anchor frame. Applying Eq. (21) would then restrict M_a to that single frame. We therefore exclude low-headroom clips when computing the restricted variant's appearance metrics. The reported comparison consequently applies only to clips for which a temporal mask can be identified.

A.9 Ablation: leave-one-out

Table 7 and Fig. 5 evaluate each term of FAM's objectives on 20 clips from EGTEA and SSv2 with a frozen TSM model. Because M_a and M_m are optimized independently, removing an appearance term leaves the motion metrics unchanged, and removing a motion term changes the appearance metrics by less than 0.005, which is within the variation between runs.

Appearance terms. Removing appearance sparsity roughly doubles the appearance mask on both datasets ($25.6\% \rightarrow 62.4\%$ on EGTEA, $32.7\% \rightarrow 65.2\%$ on SSv2) and lowers insertion ($0.812 \rightarrow 0.783$ and $0.490 \rightarrow 0.416$) while raising deletion ($0.166 \rightarrow 0.175$ and $0.077 \rightarrow 0.103$). The ℓ_1 term therefore makes the mask more compact and also improves both scores. Replacing the insertion–deletion objective with $-\log p$ produces a smaller mask on both datasets, but raises deletion from 0.166 to 0.293 on EGTEA and from 0.077 to 0.102 on SSv2, while insertion increases only on EGTEA. A mask can therefore preserve the target prediction without identifying the regions whose removal reduces it. Bilateral total variation has little effect on the numerical measures: removing it changes EGTEA insertion and deletion by less than 0.005 and increases SSv2 insertion from 0.490 to 0.503. The term controls the spatial organization of the mask rather than how much target confidence it retains, so its effect is expected to appear in the continuity and boundary alignment of the mask rather than in insertion and deletion scores.

Motion terms. Replacing headroom-normalized sufficiency with raw retention lowers necessity on both datasets, from 0.601 to 0.552 on EGTEA and from 0.727 to 0.616 on SSv2. On EGTEA, where the static reference often retains much of the target confidence, raw retention also lowers headroom-normalized sufficiency from 0.954 to 0.927 while selecting more frames ($23.1\% \rightarrow 27.5\%$). On SSv2, the static reference retains little confidence, so the two sufficiency measures nearly coincide

Data	Configuration	App Ins↑	App Del↓	Suf_{hr} ↑	Nec_{hr} ↑	$\overline{M}_a\%$	$\overline{M}_m\%$	Δcomp
EGTEA	**Full FAM**	0.812	0.166	0.954	0.601	25.6	23.1	(ref)
	− appearance ℓ_1	0.783	0.175	0.954	0.601	62.4	23.1	−0.013
	− bilateral TV	0.809	0.162	0.954	0.601	27.2	23.1	+0.000
	app. obj.: ins–del → − log p	0.833	0.293	0.954	0.601	11.8	23.1	−0.036
	− motion ℓ_1	0.812	0.168	0.954	0.581	25.6	22.5	+0.003
	mot. suf.: headroom → raw	0.813	0.165	0.927	0.552	25.6	27.5	−0.030
	prediction only	0.737	0.195	0.954	0.581	61.7	22.5	−0.031
SSv2	**Full FAM**	0.490	0.077	0.995	0.727	32.7	63.8	(ref)
	− appearance ℓ_1	0.416	0.103	0.995	0.727	65.2	63.8	−0.033
	− bilateral TV	0.503	0.078	0.995	0.727	34.0	63.8	+0.004
	app. obj.: ins–del → − log p	0.489	0.102	0.995	0.727	27.2	63.8	−0.009
	− motion ℓ_1	0.492	0.076	0.976	0.693	32.7	65.0	−0.012
	mot. suf.: headroom → raw	0.489	0.078	0.993	0.616	32.7	55.0	+0.042
	prediction only	0.432	0.126	0.976	0.693	64.9	65.0	−0.049

Table 7: **Leave-one-term ablation of FAM's objectives** on 20 clips per dataset with a frozen TSM model (single seed). Appearance metrics use the noun head on EGTEA and the action class on SSv2. $\overline{M}_a\%$ is the mean appearance-mask activation and $\overline{M}_m\%$ the mean fraction of retained frames over all clips. Δcomp is the change in $\text{mean}(\text{App Ins}, 1 - \text{App Del}, \text{Suf}_{\text{hr}}) - 0.5\,\overline{M}_m$ relative to full FAM; it penalizes motion-mask size and does not include necessity.

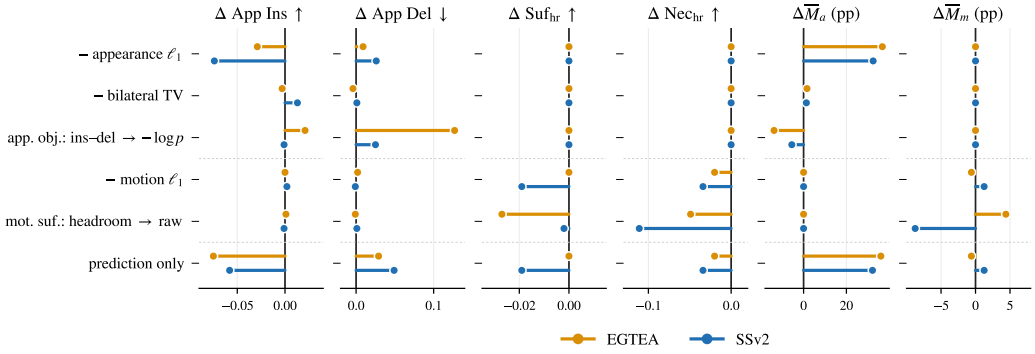


Figure 5: **Leave-one-term ablation on EGTEA and SSv2.** Each row removes or replaces one term, and each panel shows the change from full FAM in one metric (pp: percentage points). Appearance-term ablations (top) leave the motion metrics unchanged, because the two masks are optimized independently. Replacing headroom-normalized sufficiency with raw retention lowers motion necessity on both datasets.

(0.995 versus 0.993), and raw retention selects fewer frames (63.8% → 55.0%) that account for less of the model's temporal reliance. The composite score, which penalizes mask size and does not include necessity, favors raw retention on SSv2 for this reason. Removing motion sparsity has a small effect, lowering necessity by 0.020 on EGTEA and 0.034 on SSv2, and lowering SSv2 sufficiency from 0.995 to 0.976.

Regularization. Removing all regularizers leaves motion sufficiency high on both datasets (0.954 and 0.976), but the appearance mask grows to over 60%, insertion falls (0.812 → 0.737 and 0.490 → 0.432), and deletion rises (0.166 → 0.195 and 0.077 → 0.126). High sufficiency alone therefore does not imply an informative explanation.

Taken together, the insertion–deletion objective and appearance sparsity determine the quality of the appearance explanation on both datasets, and headroom normalization consistently yields motion masks with higher necessity. Bilateral total variation and motion sparsity have small effects on the reported metrics in this setting.

A.10 Ablation: causal vs. non-causal frame replacement

The main experiments use symmetric nearest-active-frame replacement from Eq. (13). Under this operator, an inactive frame can be replaced by an active frame that occurs either before or after it in the clip. We compare this choice with a causal variant that only uses active frames from the

Table 8: Causal vs. non-causal temporal replacement, TSM, N=200 clips per dataset. $\overline{A}_m$ is the mean active-frame fraction. Suf^{h} and Nec^{h} are headroom-normalized sufficiency and necessity and Suf^{raw} is the un-normalized version. Low-HR is the fraction of clips with near-zero motion headroom and is identical for both operators.

Dataset	Operator	$\overline{M}_m$	Suf^{h} ↑	Suf^{raw} ↑	Nec^{h} ↑	Low-HR
SSv2 (single-head)	current	72.5%	.992	.993	.761	8.0%
	causal-hold	78.4%	.994	.995	**.831**	
EGTEA (dual-head, verb)	current	62.3%	.939	.994	.665	70.5%
	causal-hold	61.4%	.954	.995	.656	

current or earlier time steps. The purpose of this ablation is to test whether allowing future frames as replacement sources changes the motion evaluation.

Causal replacement operator. Both variants use the same binary motion gate ${M_m}^{\text{STE}}$ from §3.1. They differ only in which active frames can replace an inactive frame.

The default symmetric operator can select any active frame:

$$w^{\text{cur}}_{j\to i} \;\propto\; \exp\!\big(-\|i-j\|^2/2\sigma_R^2\big)\, {M_m}_j^{\text{STE}}, \qquad j \neq i, \tag{22}$$

which is the weighting used in Eq. (13).

The causal variant restricts replacement sources to active frames at or before the current frame i:

$$w^{\text{causal}}_{j\to i} \;\propto\; \exp\!\big(-\|i-j\|^2/2\sigma_R^2\big)\, {M_m}_j^{\text{STE}}\, \mathbb{1}[j \le i], \qquad j \neq i, \tag{23}$$

where $\mathbb{1}[j \le i]$ is one when frame j occurs at or before frame i and zero otherwise. If no active frame occurs at or before i, the operator uses the first active frame as a fallback. This is the only case in which the causal operator uses a later frame.

For both operators, the weights are normalized and the replacement frame is selected using the same hard-argmax straight-through estimator from Eq. (A.3). The rest of the perturbation operator is unchanged.

Effect on sufficiency. The choice of replacement direction has little effect on sufficiency. On SSv2, headroom-normalized sufficiency changes from 0.992 to 0.994 and raw sufficiency changes from 0.993 to 0.995. On EGTEA, headroom-normalized sufficiency changes from 0.939 to 0.954, while raw sufficiency changes from 0.994 to 0.995. Both operators therefore preserve nearly the same amount of target confidence.

Effect on necessity. The larger difference appears in necessity on SSv2. Headroom-normalized necessity increases from 0.761 with symmetric replacement to 0.831 with causal replacement, a difference of 0.070. The active-frame fraction also increases from 72.5% to 78.4%.

One possible explanation is that symmetric replacement can use a future frame that is visually similar to the frame being removed. In the necessity test, such a replacement may preserve some of the information that would otherwise be lost when an important frame is frozen. Restricting replacements to earlier frames reduces this possibility. Under the causal operator, removing the selected frames therefore causes a larger decrease in target confidence on SSv2.

This explanation is consistent with the SSv2 label space, which contains many actions distinguished by temporal direction or order. In such cases, using content from a later point in the video at an earlier time step can preserve information about the temporal progression that the perturbation is intended to remove.

EGTEA comparison. The same change is not observed on EGTEA. Necessity changes only from 0.665 to 0.656, while headroom-normalized sufficiency increases from 0.939 to 0.954. The two operators therefore produce similar results for the motion-identifiable EGTEA clips in this experiment.

			matched budget				budget permutation		
Data	**Method**	$\lvert M_m \rvert$	AUC_k	k=2	k=4	k=6	native	permuted	Δ
EGTEA	FAM (ours)	4.6	**0.580**	**0.390**	**0.670**	0.820	0.890	0.666	**+0.224**†
	STEP-TD	3.9	0.552	0.215	0.636	**0.878**	0.592	0.571	+0.021
	EP-TD	4.6	0.555	0.261	0.658	0.876	0.706	0.705	+0.001
SSv2	FAM (ours)	5.5	**0.534**	**0.227**	**0.626**	0.823	0.902	0.696	**+0.206**†
	STEP-TD	4.2	0.480	0.156	0.529	0.768	0.524	0.526	−0.002
	EP-TD	3.7	0.527	0.191	0.575	**0.826**	0.530	0.501	+0.029

Table 9: **Budget-controlled motion faithfulness.** Left: Suf_{HR} when all methods retain the same number of frames. Right: mean Suf_{HR} at each clip's native mask size compared with the same set of budgets randomly reassigned across clips. The budget multiset is unchanged, so Δ measures whether assigning different mask sizes to different clips is informative. †None of 10,000 permutations reached the observed FAM value ($p < 10^{-4}$). The other differences are not significant after correction for multiple comparisons. AUC_k averages $k \in [1, T-1]$. For $k = 1$, $\mathrm{Suf}_{HR} = 0$ for every method because p_{static} is defined using the best repeated single frame.

EGTEA also contains many more low-headroom clips in this ablation: 70.5%, compared with 8.0% on SSv2. As a result, its normalized sufficiency and necessity values are computed from a smaller set of motion-identifiable clips. In addition, many EGTEA verbs, such as *cut*, *mix*, and *wash*, are less directly defined by temporal direction than SSv2 classes that explicitly distinguish opposite orders or directions.

A.11 Budget-controlled motion faithfulness

Motion sufficiency generally increases as more frames are retained. Since FAM can produce larger motion masks than the temporal-derived STEP and EP masks, a comparison at each method's native $|M_m|$ can mix two effects: *which* frames are selected and *how many* are retained. We separate these effects with a matched-budget evaluation.

Each method ranks frames using its own temporal score: FAM uses the soft motion mask values before binarization, while STEP-TD and EP-TD use mean at each timestep t. For a common budget k, we retain the top-k frames and apply the same nearest-active replacement operator P to all methods. We then compute

$$\mathrm{Suf}_{hr}(k) = \mathrm{clip}\left(\frac{p_v(P(x; \mathrm{top}_k)) - p_{\mathrm{static}}}{h}, 0, 1\right). \tag{24}$$

As with the other motion metrics, this analysis is restricted to headroom-identifiable clips.

FAM remains competitive at a matched budget. The large difference at each method's native mask size becomes smaller when the number of retained frames is fixed. On EGTEA, FAM improves AUC_k over STEP-TD and EP-TD by 0.028 and 0.025, respectively. On SSv2, the corresponding differences are 0.054 and 0.007. We therefore do not claim that FAM ranks frames better than the temporal-derived baselines at every budget. EP-TD slightly exceeds FAM at $k = 6$ on both datasets. One explanation is that FAM's sparsity objective is designed to identify a compact sufficient set rather than to produce a complete ranking of frames outside that set.

The largest differences occur at small budgets. FAM shows its clearest advantage when few frames are retained. At $k = 2$, FAM exceeds STEP-TD and EP-TD by 0.175 and 0.129 on EGTEA, and by 0.071 and 0.036 on SSv2. This is the regime most relevant to sparse temporal explanations, where the goal is to preserve the prediction using only a small subset of frames.

Mask size carries clip-specific information. We next test whether FAM's larger native masks alone explain its higher native sufficiency. We randomly reassign the observed mask sizes across clips while keeping the budget distribution unchanged. For FAM, this reduces mean sufficiency by 0.224 on EGTEA and 0.206 on SSv2, and none of 10,000 permutations reaches the observed value. In contrast, the corresponding changes for STEP-TD and EP-TD are small. This shows that FAM's performance is not explained only by retaining more frames: the number of retained frames is also adapted to the clip. On EGTEA, FAM and EP-TD have the same mean mask size (4.6 frames), while

Data	Clips	Operator	Δ_{sel}	Δ_{comp}	Δ_{gap}	95% CI
EGTEA	Identifiable	nearest-active	0.154	0.013	+0.141	$[+0.074, +0.215]$
		causal hold	0.171	0.028	+0.143	$[+0.082, +0.215]$
		shuffle	0.084	0.081	0.003	$[-0.027, +0.034]$
	All	nearest-active	0.081	0.009	+0.072	$[+0.049, +0.096]$
		causal hold	0.078	0.015	+0.063	$[+0.041, +0.088]$
		shuffle	0.050	0.054	-0.004	$[-0.020, +0.010]$
SSv2	Identifiable	nearest-active	0.175	-0.016	+0.191	$[+0.150, +0.237]$
		causal hold	0.208	0.105	+0.102	$[+0.046, +0.157]$
		shuffle	0.282	0.272	+0.010	$[-0.023, +0.043]$
	All	nearest-active	0.159	-0.015	+0.175	$[+0.135, +0.217]$
		causal hold	0.189	0.095	+0.094	$[+0.043, +0.142]$
		shuffle	0.267	0.255	+0.012	$[-0.017, +0.043]$

Table 10: **Frozen-mask cross-operator evaluation for FAM with TSM.** The motion mask is optimized using nearest-active replacement and evaluated without re-optimization. The primary analysis includes only headroom-identifiable clips; the all-clip results provide a robustness check. Δ_{sel} is the confidence drop when perturbing FAM-selected frames and Δ_{comp} is the drop from an equal-size subset of unselected frames. Positive Δ_{gap} indicates that the selected locations are more important than the matched complement. N varies because clips for which a valid matched comparison cannot be constructed are excluded.

their native sufficiency differs by 0.184, further showing that mean mask size alone does not explain the result.

A.12 Cross-operator evaluation of motion masks

The main FAM motion objective is optimized using nearest-active frame replacement. To test whether the learned motion mask only performs well under the same perturbation used during optimization, we freeze M_m and evaluate the selected frames using perturbations that are not used to learn the mask. Our primary analysis uses only headroom-identifiable clips ($h \geq \tau_h$), since temporal evidence cannot be reliably isolated for low-headroom predictions. We additionally evaluate all clips as a robustness check.

For each clip, we perturb matched-budget subsets of selected and unselected frames M_m and compare the resulting target-confidence drop with the drop obtained by perturbing an equal-size subset from the complement. Let Δ_{sel} and Δ_{comp} denote these drops. We report

$$\Delta_{\text{gap}} = \Delta_{\text{sel}} - \Delta_{\text{comp}}, \tag{25}$$

where a positive value indicates that the selected frames affect the prediction more than the same number of unselected frames. Nearest-active replacement is included as a reference. We use *causal hold* and *shuffle* as held-out temporal perturbations.

Held-out causal replacement. The causal-hold results are consistent across both datasets. On identifiable EGTEA clips, perturbing the selected frames produces a confidence drop of 0.171, compared with 0.028 for the matched complement, giving a gap of $+0.143$ $[+0.082, +0.215]$. On SSv2, the corresponding gap is $+0.102$ $[+0.046, +0.157]$. The same conclusion holds when all clips are included, with positive gaps of $+0.063$ on EGTEA and $+0.094$ on SSv2. Thus, the locations selected by FAM remain informative under a temporal replacement operator that is not used during optimization.

Temporal-order perturbation. Shuffle produces a different result from causal hold. On identifiable clips, the matched gaps are $+0.003$ on EGTEA $[-0.027, +0.034]$ and $+0.010$ on SSv2 $[-0.023, +0.043]$; the all-clip results are similarly near zero. Notably, shuffle causes a substantial absolute confidence drop on SSv2 for both the selected 0.282 and matched unselected 0.272 frames. Thus, SSv2 predictions are sensitive to temporal-order disruption, but the FAM-selected locations are not uniquely more order-sensitive than matched unselected locations under this perturbation.

(a) Appearance explanations

Dataset	Comparator	ΔIns ↑	p_t	ΔDel ↓	p_t	ΔM_a (pp) ↓	p_t
EGTEA	Extremal Perturbation	$+0.23^{\dagger}$	$< .001$	$-0.12^{\dagger}$	$< .001$	$-33.61^{\dagger}$	$< .001$
	STEP	$+0.16^{\dagger}$	$< .001$	$-0.18^{\dagger}$	$= .003$	$-45.34^{\dagger}$	$< .001$
SSv2	Extremal Perturbation	$+0.20^{\dagger}$	$= .003$	$-0.05^{\dagger}$	$= .013$	$-31.71^{\dagger}$	$< .001$
	STEP	$+0.17^{\dagger}$	$< .001$	$-0.07^{\dagger}$	$= .006$	$-58.53^{\dagger}$	$< .001$
	Unified Mask	$+0.19^{\dagger}$	$< .001$	$-0.19^{\dagger}$	$= .002$	$-31.35^{\dagger}$	$< .001$

(b) Motion explanations

Dataset	Comparator	$\Delta\text{Suf}_{\text{hr}}$ ↑	p_t	$\Delta\text{Suf}_{\text{raw}}$ ↑	p_t	ΔNec ↑	p_t	ΔM_m (pp) ↓	p_t
EGTEA	Extremal Perturbation–TD	$+0.20^{\dagger}$	$= .028$	$+0.08^{\dagger}$	$= .019$	$-0.08^{\dagger}$	$= .016$	$+2.41$	$= .408$
	STEP–TD	$+0.38^{\dagger}$	$= .019$	$+0.13^{\dagger}$	$= .008$	$+0.06$	$= .109$	$+10.08^{\dagger}$	$= .025$
	Motion Oracle	$-0.04^{\dagger}$	$= .043$	$+0.03$	$= .071$	$+0.14^{\dagger}$	$= .021$	$+22.49^{\dagger}$	$= .009$
SSv2	Extremal Perturbation–TD	$+0.53^{\dagger}$	$= .004$	$+0.47^{\dagger}$	$= .004$	$+0.16^{\dagger}$	$= .019$	$+18.98^{\dagger}$	$= .002$
	STEP–TD	$+0.51^{\dagger}$	$< .001$	$+0.46^{\dagger}$	$< .001$	$+0.20^{\dagger}$	$= .022$	$+15.64^{\dagger}$	$= .004$
	Motion Oracle	$+0.00$	$= .202$	$+0.00^{\dagger}$	$= .008$	$+0.20^{\dagger}$	$= .008$	$+25.85^{\dagger}$	$< .001$

Table 11: **Paired statistical comparison of FAM with appearance and motion baselines on EGTEA and SSv2 using TSM.** Each entry reports the mean paired difference $\Delta = \text{FAM} - \text{comparator}$ across three seeds. Positive differences favor FAM for insertion, sufficiency, and necessity, whereas negative differences favor FAM for deletion and mask area. Mask-area differences are reported in percentage points (pp). p_t denotes the uncorrected two-sided paired t-test p-value with $n = 3$ paired observations and 2 degrees of freedom. † denotes $p_t < .05$. The exact motion oracle is included only as a reference and is not a practical explanation method. Wilcoxon signed-rank results are omitted because, with three paired observations, the smallest attainable nonzero two-sided p-value is 0.25.

Overall, FAM-selected locations remain more influential than matched unselected locations under held-out causal replacement, showing that their importance is not specific to the nearest-active replacement used during optimization. This advantage does not extend to shuffle, which tests sensitivity to ordering rather than replacement of temporal evidence.

A.13 Statistical Significance

Statistical comparison. Table 11 shows that FAM improves appearance insertion, deletion, and spatial-mask size relative to every appearance comparator on both datasets. On EGTEA, insertion increases by 0.16–0.23, deletion decreases by 0.12–0.18, and the appearance mask is reduced by 33.61–45.34 percentage points relative to Extremal Perturbation and STEP. On SSv2, insertion increases by 0.17–0.20, deletion decreases by 0.05–0.19, and the appearance mask is reduced by 31.35–58.53 percentage points relative to Extremal Perturbation, STEP, and the Unified Mask variant.

On the motion axis, FAM obtains higher headroom-normalized and raw sufficiency than both temporal baselines on both datasets. On EGTEA, headroom-normalized sufficiency increases by 0.20 relative to Extremal Perturbation–TD and 0.38 relative to STEP–TD. The improvement is larger on SSv2, where headroom-normalized sufficiency increases by 0.53 and 0.51, and raw sufficiency by 0.47 and 0.46, respectively. Necessity gives a more qualified result. On SSv2, FAM obtains higher necessity than both temporal baselines ($+0.16$ and $+0.20$). On EGTEA, Extremal Perturbation–TD obtains higher necessity than FAM (-0.08), and the difference from STEP–TD ($+0.06$) is not significant.

FAM's higher sufficiency on SSv2 comes with larger temporal masks: FAM retains 18.98 percentage points more frames than Extremal Perturbation–TD and 15.64 more than STEP–TD. On EGTEA, the difference from Extremal Perturbation–TD ($+2.41$) is not significant, while FAM retains 10.08 percentage points more frames than STEP–TD. FAM therefore preserves substantially more motion-dependent confidence, but does not produce the smallest temporal explanation.

Compared with the exact motion oracle, FAM retains 22.49 percentage points more frames on EGTEA and 25.85 on SSv2, and obtains higher necessity on both datasets ($+0.14$ and $+0.20$). On

EGTEA, the oracle's headroom-normalized sufficiency is slightly higher (-0.04). On SSv2, the difference in headroom-normalized sufficiency is not significant, and although the rounded raw-sufficiency difference is 0.00, the unrounded difference is positive and significant. FAM therefore reaches similar sufficiency to the oracle while retaining more frames than the smallest subset found through exhaustive search. Its higher necessity is consistent with its larger masks, since freezing more frames removes more confidence.

A.14 Sensitivity Analysis

Figure 6 shows a plot of the composite value for different values of λ_{BTV} and λ_M. The winning configuration is marked with a star. For λ_{BTV}, we see a clear front of values leading to higher composite values, while λ_M has multiple explanations of different sizes that still pass the sufficiency criteria.

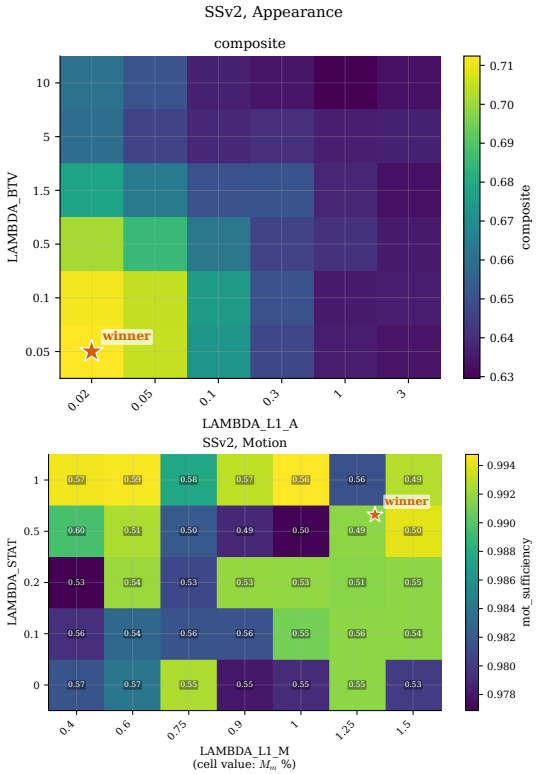


Figure 6: Hyperparameters sensitivity for SSV2 dataset.

A.15 Reproducibility Statement

Code, the grid search artifacts used to set $(\lambda_A, \lambda_M, \lambda_{BTV}, \lambda_{stat})$ (Appendix Table 12), and trained checkpoints will be released upon acceptance. Frozen-classifier weights are taken from the official releases of the cited backbones for SSv2. The TimeSformer and SlowFast models trained on EGTEA will be released upon acceptance.

A.16 Hyperparameters

Table 12 provides the full list of hyperparameters used to obtain the results presented in this paper.

A.17 AI Usage Disclosure

Generative AI tools were used during manuscript preparation to assist with only language editing. They were also used as a programming aid for debugging and drafting portions of experimental code. All generated text and code suggestions were reviewed, modified where necessary, and verified by

Table 12: **FAM hyperparameters for EGTEA and SSV2 dataset.** Settings are identical for EGTEA Gaze+ and Something-Something V2 except the λ parameters.

Hyperparameter	Value	Description
Iterations	500	M_m and M_a mask optimization steps
Learning rate	0.05	Adam, both stages
M_m LR decay	final 20%	Linear decay to zero
Frames T	8	Uniformly sampled frames per clip
Resolution	224×224	Input spatial resolution
λ_{pred}	1.0	Appearance insertion–deletion weight
λ_{suf}	3.0	Motion sufficiency weight
λ_M	2.0/1.25	Motion-mask ℓ_1 sparsity for EGTEA/SSV2
λ_A	0.02	Appearance-mask ℓ_1 sparsity
λ_{BTV}	0.05	Bilateral-TV regularization
λ_{stat}	1.5	Static prior for SSV2
τ_s	0.95	Headroom-normalized sufficiency threshold
β_s	0.05	Sufficiency softplus smoothing
τ_h	0.05	Motion-headroom threshold
σ_R	2.0	Temporal replacement bandwidth
Motion initialization	0.5 ± 0.05	Initial motion-mask logits
Appearance initialization	0.7	Initial appearance-mask seed
Appearance scale r	7	Spatial seed downsampling factor
Blur (k_B, σ_B)	$(31, 20)$	Appearance replacement blur

the authors. The authors designed the methods, experiments, executed all experiments, analyzed the reported results, and are responsible for all scientific claims and conclusions in the paper.