Disentangling Where and When: Factored Spatio-Temporal Explanations for Video Action Recognition

Bhavan Vasu
Oregon State University
vasub@oregonstate.edu

Giuseppe Raffa
Intel Labs
giuseppe.raffa@intel.com

Prasad Tadepalli
Oregon State University
prasad.tadepalli@oregonstate.edu

Abstract

Post-hoc explanation methods for video action recognition typically conflate spatial appearance and temporal dynamics into a single saliency or importance map. We propose a perturbation-based framework that disentangles these contributions into a spatial appearance mask (M_a) and a temporal motion mask (M_m). The two masks are jointly optimized against a frozen action recognition model through factored perturbation architecture, with nearest frame replacement to identify important temporal dynamics. A separation penalty discourages the appearance mask from spending budget on frames that the motion mask has already discarded. A ReLU-based area constraint penalizes explanations larger than an assigned budget. To evaluate these explanations, we introduce a separation protocol that sweeps spatial pixels and temporal frames independently, backed by proofs characterizing the proposed method. Across EPIC-Kitchens-55, EGTEA Gaze+, and Something-Something V2, our method attains the lowest deletion AUC against leading baselines while producing masks roughly five times sparser.

1 Introduction

Video action recognition models such as TSM [16], SlowFast [4], and Video Swin [17] now reach high accuracy on standard benchmarks. They increasingly inform downstream decisions in settings where errors are costly, such as assistive robotics in unstructured home environments [2], surgical workflow analysis and skill assessment [22, 6]. In each of these settings, top-1 accuracy on a held-out test split is a necessary but insufficient guarantee. A practitioner who deploys a video classifier needs to know *why* the model made a given prediction before deciding whether to trust it, deploy it, or audit it for bias.

For static image classification, this question has been studied extensively. Gradient-based methods [20, 23] and perturbation-based methods [18, 5] produce spatial saliency maps that highlight image regions correlated with the predicted class. The transition to video introduces a complication that these methods do not address. A video classifier is conditioned on two qualitatively different sources of evidence: the *appearance* of objects within frames, and the *temporal dynamics* across frames. The practical consequences of relying on one source rather than the other are not interchangeable.

Motivation: Consider a kitchen-assistance robot that uses a video classifier to recognize the action "put-down plate" from a hand-mounted camera as shown in Figure 1. There are two qualitatively different ways the classifier might be right. It might recognize the plate (an appearance cue: the noun) and infer the action from context. It might instead recognize the downward trajectory (a motion cue:

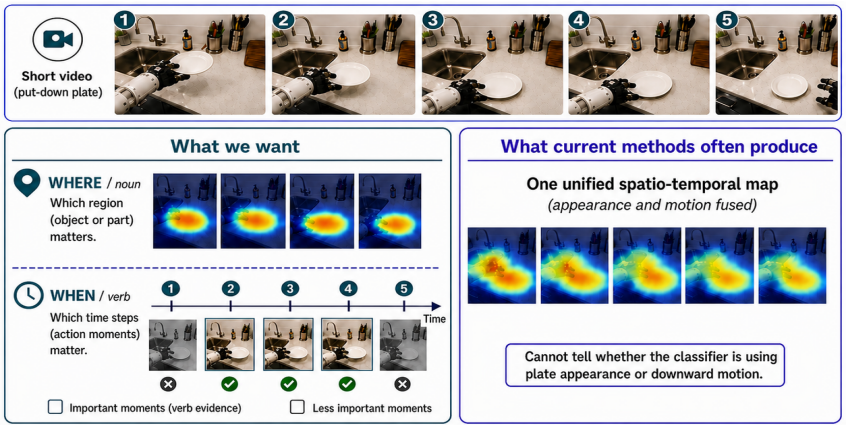


Figure 1: Factored explanations separate appearance (noun: plate) evidence, where the model looks , from motion evidence (verb: put-down), when the model acts. Standard saliency instead yields a single unified map that conflates appearance and motion.

the verb) and infer the object from context. A standard saliency map cannot distinguish these two cases, but their failure modes differ sharply. A model that keys primarily on plate appearance will misfire when the plate is occluded by the user's hand or replaced by a visually similar distractor (e.g., a flat lid). A model that keys primarily on the downward motion will misfire when the user lowers something that is not a plate, or when the plate is set down in a way that does not match the typical trajectory. The same accuracy figure on a balanced test set hides both failure modes; a factored explanation surfaces them.

Applications: Disentangling appearance and motion enables critical capabilities across domains. Factored explanations support shortcut auditing by separating spurious appearance correlations (e.g., background bias) from spurious motion correlations (e.g., generic hand sweeps), which single-mask methods conflate. This distinction also supports robustness forecasting: knowing whether a model relies on appearance (vulnerable to occlusion) or motion (vulnerable to viewpoint and framerate) informs deployment constraints and data augmentation. In clinical video analysis, it separates reliance on instrument appearance from procedure-specific motion to help align predictions with expert protocols. In egocentric assistive technology, it distinguishes active user intent ("reaching for X") from passive context ("X is present"). At the dataset level, aggregating masks can reveal static shortcuts, visual ambiguity, and alignment with human attention datasets like EGTEA Gaze+ [14] by matching spatial masks to gaze location and temporal masks to fixation timing.

Why existing methods fall short. Gradient-based methods such as GradCAM [20] and Score-CAM [23], when applied per-frame to a video model, produce spatial heatmaps that conflate spatial and temporal contributions. Aggregating per-frame gradient magnitudes into a temporal importance signal (Temporal GradCAM) preserves time but discards space. Perturbation-based methods adapted to video [18, 5, 15] optimize a single spatiotemporal mask, which suffers from the same conflation: the mask is high in regions where *either* the spatial pixel *or* the temporal frame matters, and the user cannot read off which. Unlike the recent DANCE framework [13] that resolves the conflation through an ante-hoc concept bottleneck, our work targets the post-hoc, vocabulary-free regime: factored explanations for any frozen classifier expressed as two coupled but semantically distinct masks. The framework is built around four design choices:

1. **Factored perturbation via Straight-Through Estimator (STE)** (§3.1). We disentangle evidence into a spatial appearance mask (pixel-level blur) and a temporal motion mask (discrete frame replacement) using the STE.
2. **One-sided area constraint** (§3.2). We replace standard MSE area penalties with a ReLU upper bound, which allows the learned masks to settle at extreme sparsity levels well below the budget.
3. **Separation loss for optimization degeneracy** (§3.2). We introduce a strict penalty that forces the appearance mask to zero on temporally frozen frames.

4. **Semantic conditioning and flow alignment** (§3.2). We structurally route gradients by supervising the appearance mask with noun confidence and the motion mask with verb confidence. For models lacking dual heads, we substitute semantic conditioning with optical flow alignment.

Standard insertion/deletion metrics [18] evaluate a single mask against a single class probability. Instead, we introduce a **separation evaluation protocol** (§4) that sweeps each axis independently: appearance sweeps track *noun* confidence, while motion sweeps track *verb* confidence. We evaluate on EPIC-Kitchens-55 [2], EGTEA Gaze+ [14], and Something-Something V2 [7] against STEP [15], Extremal Perturbation [5], and a unified-mask ablation. Our method achieves the lowest deletion AUC across most datasets and axes, producing masks approximately $5\times$ sparser than the strongest baseline. While Insertion AUC inherently favors the larger masks of our baselines (an artifact of rewarding absolute pixel count, detailed in §5), our method attains the highest motion complementary fidelity (F^-) across all datasets, directly validating our core disentanglement claim.

2 Related work: Spatio-temporal explanations for video

Temporal GradCAM aggregates per-frame gradient magnitudes into a one-dimensional importance signal, preserving time at the cost of spatial resolution. SWAG-V [8] computes superpixel-weighted gradient maps over video, producing piecewise-constant attributions that are easier to read than dense heatmaps but inherit gradient methods' sensitivity to model nonlinearities. Saliency Tubes [21] reshape the activations of a video network's last 3D convolutional layer into volumetric heatmaps. None of these distinguish spatial from temporal contributions in the resulting map. Perturbation-based methods treat the classifier as a black box and observe confidence change under input modification. STEP [15] optimizes a per-frame spatial mask under a Gaussian-blur baseline with an MSE area loss for sparsity. Extremal Perturbation [5] was originally proposed for images and extends to video by jointly optimizing a smooth fixed-area mask over the spatio-temporal volume. We use STEP and Extremal Perturbation as perturbation-based baselines in Section 5; both return a single mask in which spatial and temporal evidence are entangled.

Concept-based methods describe a prediction through higher-level features rather than raw pixels. VTCD [12] discovers reusable spatio-temporal concept tubes via clustering and quantifies their effect on transformer predictions. Ji et al. [10] extend TCAV to 3D ConvNets, and Saha et al. [19] report that dynamic spatio-temporal concepts carry more information than static ones for action classification. MoTIF [11] extends concept bottlenecks to video with per-concept temporal self-attention, explaining *when* discovered object and action concepts contribute to a prediction. These approaches test whether a concept influences a prediction without localizing the influence to specific pixels or frames. Language-based variants [24] target semantic-level interpretability through prior knowledge and do not address the appearance motion distinction directly. Contemporaneous to our work, DANCE [13] retrains classifiers with a concept bottleneck that separates motion (2D pose clusters) from spatial context (LLM-proposed objects) to attribute predictions to discrete named concepts. Unlike DANCE [13], which retrains models to extract discrete, vocabulary-constrained spatial and motion concepts, our method is *post-hoc* (requires no retraining), outputs *continuous* pixel/frame masks for precise localization, imposes *no vocabulary limits* (handling non-human or occluded actions), and evaluates *empirical* rather than architectural faithfulness.

3 Method

Let $f : \mathbb{R}^{3\times T\times H\times W} \to \mathbb{R}^{|\mathcal{V}|} \times \mathbb{R}^{|\mathcal{N}|}$ be a frozen dual-head video classifier producing verb logits $\mathbf{z}^v$ and noun logits $\mathbf{z}^n$ from a clip $\mathbf{x}$ of T uniformly-sampled RGB frames at $H \times W$ resolution. Let $v^* = \arg\max_v \text{softmax}(\mathbf{z}^v(\mathbf{x}))_v$ and $n^* = \arg\max_n \text{softmax}(\mathbf{z}^n(\mathbf{x}))_n$ denote the predicted verb and noun on the clean clip. We seek two masks $M_a \in [0,1]^{T\times H\times W}$(appearance: where) and $M_m \in [0,1]^T$(motion: when), such that:

$$\begin{aligned} \max_{M_a, M_m} \quad & \text{softmax}(\mathbf{z}^v(\hat{\mathbf{x}}))_{v^*} + \text{softmax}(\mathbf{z}^n(\hat{\mathbf{x}}))_{n^*} \\ \text{s.t.} \quad & \frac{1}{THW}\sum M_a \le B_a, \quad \frac{1}{T}\sum M_m \le B_m \end{aligned} \tag{1}$$

Where M_a identifies the spatial pixels critical for n^* and M_m identifies the temporal frames critical for v^*, $B_a, B_m \in (0,1)$ define the maximum allowable spatial and temporal budgets, forcing the masks

to localize minimal subsets of critical evidence. M_a marks the pixels whose preservation (versus a heavy blur) sustains the noun prediction, and M_m marks the frames whose preservation (versus nearest-active-frame replacement) sustains the verb prediction. For single-head models that emit only a single logit vector $\mathbf{z} \in \mathbb{R}^{|\mathcal{C}|}$, the same mask shapes are retained and the noun/verb decomposition is replaced by the optical-flow alignment loss of §3.3. Both masks are parameterized as raw logits and converted to mask values through a temperature-annealed sigmoid $\sigma_\tau(x) = 1/(1 + e^{-\tau x})$, with τ rising linearly from τ_{start}=1 to τ_{max}=10 over the course of optimization. A low temperature in early iterations preserves gradient flow; a high temperature in later iterations sharpens the mask toward $\{0, 1\}$. The appearance mask is optimized at low resolution $H_s = W_s = \lfloor H/r \rfloor$ with $r = 7$, then bilinearly upsampled to $H \times W$, Gaussian-smoothed (kernel size $2r - 1$, $\sigma = r/2$), and clamped to $[0, 1]$. The motion mask is parameterized as T scalar logits, one per frame.

3.1 Factored perturbation with hard frame selection

The perturbation is a single compositional forward pass with two factors: a spatial blur based blend governed by the appearance mask M_a, and a temporal nearest-frame replacement governed by the motion mask M_m that consumes the spatially-perturbed clip as its replacement source. Both masks are parameterized by logits (ℓ_A, ℓ_M) and optimized *jointly* under a single objective (Sec. 3.2); the "factor" ordering below describes the forward composition, not a staged optimization schedule.

Spatial factor: blur–blend. Given the clean clip $\mathbf{x}$ and a heavily blurred copy $\tilde{\mathbf{x}} = \mathcal{G}_B * \mathbf{x}$ (kernel size k_B=31, σ_B=20),

$$\mathbf{x}^{\text{app}} = M_a \odot \mathbf{x} + (1 - M_a) \odot \tilde{\mathbf{x}}. \tag{2}$$

Blur is preferred to a zero or mean-frame baseline because it preserves low-frequency context, keeping the perturbed clip closer to the model's training distribution and reducing the risk that observed confidence drops are attributable to distribution shift rather than removed evidence.

Temporal factor: nearest-frame replacement. Frames that the motion mask designates as inactive are replaced by the content of the nearest *active* frame from $\mathbf{x}^{\text{app}}$. The replacement source is the spatially-perturbed clip rather than the original, so the appearance mask of the source frame is propagated to its replacement targets; the choice of replacement source is discrete in the forward pass.

Concretely, let $M_m = (m_1, \ldots, m_T)$. The proximity-weighted, motion-gated replacement weights

$$w_{i \to j} = \frac{\exp(-\|i - j\|^2 / 2\sigma_R^2) \cdot m_j}{\sum_{k \neq i} \exp(-\|i - k\|^2 / 2\sigma_R^2) \cdot m_k + \epsilon}, \qquad \sigma_R = 2. \tag{3}$$

A pure soft average over $\{w_{i \to j}\}$ produces a blurred replacement frame, reintroducing the artifact the temporal factor was intended to avoid. We instead take the argmax in the forward pass and let gradients flow through the soft weights via a Straight-Through Estimator:

$$j^*(i) = \arg\max_j w_{i \to j}, \qquad \tilde{w}_{i \to j} = \mathbb{1}[j = j^*(i)] - \text{sg}(w_{i \to j}) + w_{i \to j}. \tag{4}$$

A second STE binarizes the temporal gate itself,

$$M_{m_t}^{\text{hard}} = \mathbb{1}[m_t > 0.5], \qquad M_{m_t}^{\text{STE}} = M_{m_t}^{\text{hard}} - \text{sg}(m_t) + m_t, \tag{5}$$

and the final perturbed clip is

$$\hat{\mathbf{x}}_t = M_{m_t}^{\text{STE}} \cdot \mathbf{x}_t^{\text{app}} + (1 - M_{m_t}^{\text{STE}}) \cdot \sum_j \tilde{w}_{t \to j} \, \mathbf{x}_j^{\text{app}}. \tag{6}$$

At convergence, frozen frames duplicate the nearest active frame to structurally remove temporal dynamics. Preserving gradient flow via a Straight-Through Estimator (STE) introduces a visual discrepancy between the true discrete selection ($\hat{\mathbf{x}}_i^{\text{hard}}$) and its continuous proxy ($\hat{\mathbf{x}}_i^{\text{soft}}$).

Proposition 1 (STE bias bound)**.** If Δ_i is the probability gap between the top two replacement candidates, and let δ_i is the maximum visual discrepancy among candidates. Then, the pointwise discrepancy between $\hat{\mathbf{x}}_i^{\text{hard}}$ and $\hat{\mathbf{x}}_i^{\text{soft}}$ is bounded by (proof in App. B.1) :

$$\|\hat{\mathbf{x}}_i^{\text{hard}} - \hat{\mathbf{x}}_i^{\text{soft}}\| \;\leq\; (1 - \Delta_i)\delta_i$$

Empirically, temperature annealing drives the mask toward a discrete one-hot state ($\Delta_i \to 1$), safely collapsing this discrepancy to zero. Choosing $\mathbf{x}^{\text{app}}$ as the replacement tightens the STE bias bound of Proposition 1: the visual discrepancy δ_i between candidate replacement frames is smaller for spatially-cleaned frames than for raw frames, since M_a removes irrelevant background variation, so the bound $(1 - \Delta_i)\delta_i$ contracts.

Alternative strategies for replacing temporally inactive frames introduce severe artifacts: zero or mean baselines cause distribution shift, frame shuffling creates spurious trajectories (failing on circular motions), and soft blending leaves residual ghosting. By replacing inactive frames with exact copies of their nearest active neighbors, the method creates static blocks of identical frames. This completely halts the motion trajectory, reducing temporal dynamics to exactly zero rather than merely attenuating them, while preserving spatial appearance. Crucially, this nearest-neighbor approach minimizes the visual discrepancy factor δ_i, directly tightening the formal STE bias bound (Proposition 1).

3.2 Loss functions

For dual-head models, the total loss is

$$\mathcal{L} = \underbrace{-\log p_{n^*}(\hat{\mathbf{x}}) - \log p_{v^*}(\hat{\mathbf{x}})}_{\text{split prediction}} + \underbrace{\lambda_A \mathcal{L}^a_{\text{area}} + \lambda_M \mathcal{L}^m_{\text{area}}}_{\text{sparsity}} + \underbrace{\lambda_S \mathcal{L}_K}_{\text{smoothness}} + \underbrace{\lambda_D \mathcal{L}_{\text{sep}}}_{\text{separation}}. \tag{7}$$

Split prediction. The noun log-probability $-\log(\text{softmax}(\mathbf{z}^n(\hat{\mathbf{x}}))_{n^*})$ supervises the appearance mask, and the verb log-probability supervises the motion mask. The logarithmic form is preferred to the linear form $-p$ because the gradient of $-\log p$ does not vanish as $p \to 0^+$, which accelerates convergence in early iterations when the masks are nearly empty and target-class probability is consequently low. Single-head models add the flow alignment loss of §3.3, and replace the split prediction with a single target-class log-probability.

Measured at convergence, cross-axis gradients are of the same order as the intended ones (median leak ratio 0.85 for M_a and 1.23 for M_m; App. B.5). Semantic separation of the explanations is therefore an empirical outcome that depends on the classifier's learned representations and on the available supervision; removing the noun/verb split degrades deletion AUC and F^- on both axes (App. B.4).

Sparsity constraint. Prior work [15] realizes the sparsity constraint in Equation 1 through an MSE loss that penalizes the sorted mask values for deviating from a binary target with B_a or B_m active entries. This penalty is symmetric: the mask is penalized for using *fewer* pixels than the budget as well as more, which forces the optimizer to occupy the budget exactly regardless of input complexity. We argue that an explanation should be as sparse as possible within the budget. Replacing the MSE with a ReLU upper bound,

$$\mathcal{L}^{\text{ReLU}}_{\text{area}}(\mathbf{m}, v) = \frac{1}{N} \sum_{i=1}^{N} \left[\text{ReLU}(\text{sort}(\mathbf{m})_i - t_i)\right]^2, \tag{8}$$

penalizes only excess. Empirically, the realized appearance mask settles at roughly $5\times$ smaller than the budget on EPIC-Kitchens-55 and EGTEA (Table 1), whereas under the symmetric MSE form the mask occupies the budget exactly. Appendix B.2 formalizes this mathematical distinction: under standard weak monotonicity assumptions on the prediction loss landscape, the symmetric MSE form provably forces the realized area to saturate the budget at a rate of $\mathcal{O}(1/\lambda)$ (Proposition 2).

Proposition 2 (MSE forces budget saturation, with quantitative rate)**.** Let $\mathbf{m}^*_\lambda$ be an interior stationary point of $\mathcal{L}_{\text{pred}} + \lambda \mathcal{L}^{\text{MSE}}_{\text{area}}$. As $\lambda \to \infty$, the realized area a^*_λ converges to the target budget b, and the gap closes at a rate of $O(1/\lambda)$ with a constant scaling of $\sqrt{N}$. Specifically, the realized area satisfies the following bound:

$$|a^*_\lambda - b| \;\le\; \frac{L\sqrt{N}}{2\lambda}$$

Conversely, the ReLU formulation analytically permits sub-budget stationary points, allowing the realized sparsity to be dictated entirely by the underlying prediction-loss geometry independent of the penalty weight λ (proof App. B.2).

Proposition 3 (ReLU permits sub-budget stationary points independent of λ)**.** Suppose there exists a mask $\mathbf{m}^* \in [0,1]^N$ that is a stationary point of the prediction loss ($\nabla \mathcal{L}_{\text{pred}}(\mathbf{m}^*) = \mathbf{0}$) and naturally satisfies the upper-bound budget constraint ($\mathbf{m}_i^{\text{sort},*} \leq t_i$ for all i, yielding a realized area $a^* \leq b$). Then, $\mathbf{m}^*$ remains a valid stationary point of the joint loss for every penalty weight $\lambda > 0$, satisfying:

$$\nabla\Big(\mathcal{L}_{\text{pred}}(\mathbf{m}^*) + \lambda \mathcal{L}_{\text{area}}^{\text{ReLU}}(\mathbf{m}^*, v)\Big) = \mathbf{0}$$

Ellipsoid-kernel smoothness. Following [15], we convolve M_a with a 3D ellipsoid kernel (t_k=3, h_k=w_k=31, supported on the unit ball) and apply the same sorted-target loss as the sparsity term. This regularizer encourages the mask to form connected spatiotemporal regions rather than isolated high-activation pixels.

Separation. The separation loss penalizes appearance activation in temporally frozen frames:

$$\mathcal{L}_{\text{sep}} = \frac{1}{THW} \sum_{t,h,w} M_{a,t,h,w} \cdot (1 - M_{m,t}). \tag{9}$$

The motivation is precise: when $M_{m,t} \approx 0$, frame t is replaced by content from a nearby active frame, so any spatial highlights M_a assigns to the original frame t are never observed by the model. Without an explicit penalty, the optimizer is free to expend appearance budget on these unobserved positions. With Eq. (9), the appearance mask concentrates on the frames that survive temporal gating.

3.3 Flow alignment for single-head models

Something-Something V2 contains 174 composite action classes that admit no natural verb/noun decomposition (e.g., *pushing something from left to right* versus *pushing something from right to left*). The split prediction loss therefore has nothing to split. To drive M_a and M_m toward semantically distinct quantities, we anchor the M_m to a physics-grounded signal: per-frame optical flow magnitude.

Let $\bar{\phi}_t = \|\text{FlowMag}(t)\| / \max_t \|\text{FlowMag}(t)\| \in [0,1]$ denote the normalized flow magnitude at frame t, computed via [3]. The flow alignment loss is

$$\mathcal{L}_{\text{flow}} = 1 - \cos(M_m, \bar{\phi}). \tag{10}$$

This biases M_m toward frames in which pixels are observably moving, while leaving the prediction loss (a single $-\log p_{c^*}$) free to drive the appearance mask. We acknowledge this represents a pragmatic empirical heuristic rather than a structurally guaranteed decoupling; however, anchoring the temporal axis to explicit optical flow provides the necessary inductive bias to separate temporal dynamics from static spatial features when independent semantic verb/noun logits are unavailable. Flow alignment is a real inductive bias rather than a neutral regularizer. The learned M_m is only moderately correlated with flow magnitude ($r = 0.37$), so it is influenced by, but does not reproduce, the flow signal. Our strongest appearance/motion separation results are therefore on the dual-head datasets, where the separation is driven by the model's own noun and verb outputs.

4 Evaluation protocol

Standard insertion/deletion metrics [18] sweep a single mask in order of importance and track one class probability. For a factored explanation, this answers the wrong question: it reports how much confidence is recovered from the union of spatial and temporal cues, not how much each axis contributes. We propose a separation protocol that treats the two axes independently.

Appearance axis. Sort spatial positions in descending order of M_a, with frames outside M_m kept frozen. For a fraction α, keep the top-α positions sharp and blur the rest, and track the noun (dual-head) or class (single-head) confidence:

$$\text{Ins}_a(\alpha) = \text{softmax}\Big(\mathbf{z}^n\Big(M_a{}^{(\alpha)} \odot \mathbf{x} + (1 - M_a{}^{(\alpha)}) \odot \tilde{\mathbf{x}}\Big)\Big)_{n^*}. \tag{11}$$

Deletion is defined symmetrically using the complement mask. The area under the insertion curve $\text{AUC}_{\text{Ins}}^a = \int_0^1 \text{Ins}_a(\alpha)\, d\alpha$ is to be maximized; the area under the deletion curve is to be minimized.

Table 1: Explanation quality comparison. Each cell reports TSM/TimeSformer. App/Mot denote appearance/motion; Ins/Del denote insertion/deletion AUC. F^+ and F^- are fidelity metrics; M_a, M_m are active pixel/frame fractions. Color intensity within each dataset/model column maps to relative performance (darker = better, computed by min–max normalization, bucketed into 6 levels). Paired significance tests and standard deviations are reported in Appendix B.7.

	Insertion AUC		**Deletion AUC**		**Fidelity+**		**Fidelity−**		**Mask size**	
Method	App↑	Mot↑	App↓	Mot↓	App↑	Mot↑	App↑	Mot↑	M_a ↓	M_m ↓
EGTEA (N = 400), TSM/TimeSformer										
EP + M_m	.72/.65	.79/.71	.41/.24	.77/.67	.87/.85	.87/.80	.14/.40	.11/.17	50.5/50.7	61.6/61.6
EP	.70/.65	–	.43/.27	–	.91/.96	–	.12/.33	–	50.6/50.8	–
STEP + M_m	.77/.72	.84/.82	.46/.37	.79/.74	.97/.97	.96/.96	.43/.72	.12/.20	60.6/69.0	81.8/83.3
STEP	.70/.73	–	.51/.41	–	.99/.99	–	.28/.57	–	56.4/65.4	–
Unified Mask	.81/.78	–	.43/.36	–	1.00/1.00	–	.15/.23	–	20.2/18.2	–
Ours (Full)	.75/.63	.79/.88	.27/.20	.25/.35	.99/.99	.99/.99	.25/.41	.82/.65	12.4/10.7	37.9/33.6
EPIC-Kitchen-55 (N = 9385), TSM/TimeSformer										
EP + M_m	.40/.54	.53/.39	.15/.12	.50/.37	.54/.41	.33/.43	.57/.23	.20/.08	50/50.3	61.5/61.4
EP	.41/.31	–	.18/.14	–	.58/.54	–	.30/.22	–	50.4/49.2	–
STEP + M_m	.58/.45	.70/.61	.20/.18	.62/.54	.89/.80	.65/.75	.87/.42	.21/.02	70.2/75.2	89.8/86.8
STEP	.57/.43	–	.23/.17	–	.94/.88	–	.55/.37	–	66.1/73.3	–
Unified Mask	.59/.39	–	.22/.16	–	.99/.99	–	.28/.18	–	21.1/19.6	–
Ours (Full)	.44/.29	.66/.77	.10/.08	.21/.25	.97/.96	.98/.96	.44/.26	.62/.29	10.3/9.6	24.5/27.6
SSv2 (N = 200), TSM/TimeSformer										
EP + M_m	.26/.42	.40/.47	.08/.13	.39/.47	.53/.62	.55/.62	.40/.39	.47/.46	52.0/51.9	64.4/64.4
EP	.32/.48	–	.14/.18	–	.83/.84	–	.28/.29	–	52.0/51.9	–
STEP + M_m	.36/.56	.48/.58	.15/.22	.44/.56	.93/.95	.93/.94	.66/.68	.61/.48	79.3/78.8	86.7/85.5
STEP	.34/.52	–	.17/.24	–	.99/.99	–	.64/.60	–	78.9/78.7	–
Unified Mask	.39/.47	–	.18/.29	–	1.00/1.00	–	.63/.54	–	51.1/51.6	–
Ours (Full)	.31/.37	.25/.46	.11/.10	.11/.32	.93/.90	.13/.41	.60/.52	.62/.42	22.8/24.0	22.1/26.2

Motion axis. Sort frames in descending order of M_m. Starting from the M_a-masked clip, un-freeze (insertion) or freeze (deletion) frames in that order using the nearest-active-frame replacement, and track the verb or class confidence. The two AUCs $\text{AUC}^m_{\text{Ins}}$ and $\text{AUC}^m_{\text{Del}}$ are interpreted as on the appearance axis.

Fidelity. F^+ is the fraction of clips for which the masked input preserves the model's original top-1 prediction; F^- is the fraction for which the *complement* of the masked input changes it. Both are to be maximized, per axis. F^+ asks whether the mask retains the evidence; F^- asks whether the mask retains *only* the evidence. The latter is the more direct test of explanation faithfulness, since recovering confidence with a large mask is easier than destroying it with a small one.

Sparsity. $\overline{A}_a$ and $\overline{A}_m$ denote the mean active fractions of pixels and frames, averaged over the test set. Because Eq. (8) permits below-budget operation, $\overline{A}$ may fall well below the optimization target. Standard deviations are reported as well: high variance is itself diagnostic, indicating that some action classes require substantially more evidence than others.

5 Experiments

5.1 Setup

Datasets and Models We evaluate on three benchmarks. EPIC-Kitchens-55 [2] provides egocentric kitchen video with $(verb, noun)$ pair labels and serves as our primary dual-head testbed.

EGTEA Gaze+ [14] provides a smaller egocentric corpus with synchronized gaze recordings with $(verb, noun)$ pair labels. Something-Something V2 [7] contains 174 single-label composite classes whose discrimination requires temporal reasoning, and is the testbed for the flow-alignment variant. Table 1 presents results on a randomly picked subset (listed in table) of the three datasets. For EPIC-Kitchens-55 and EGTEA Gaze+ we use frozen TSM [16] (ResNet-50) and TimeSformer [1] models with separate verb and noun heads and for Something-Something V2 we use the single-head versions of the same backbones.

5.2 Qualitative results

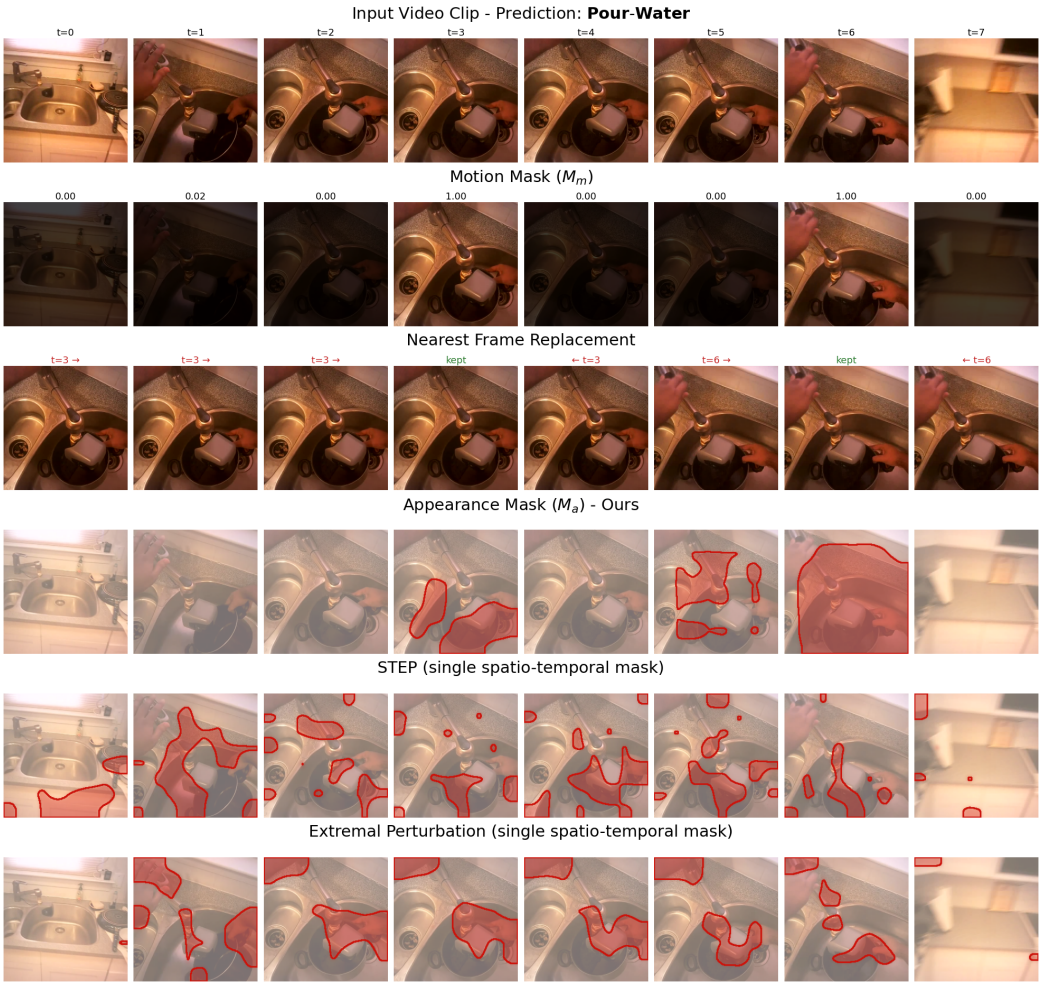


Figure 2: Qualitative figure showing (a) Input video clip; (b) Motion mask M_m; (c) The video clip with the winning nearest-frame replacement that resulted in M_m; (d) Appearance mask M_a produced by our approach, STEP and EP. Highlighted regions in M_a show the top 15% of important regions.

Baselines. Extremal Perturbation [5], STEP [15] and a Unified Mask ablation that optimizes a single $M_a{}' \in [0,1]^{T\times H\times W}$ within our perturbation pipeline but without the appearance/motion factorization. The unified ablation isolates the contribution of the factored design itself, controlling for the perturbation baseline, optimization schedule, and area-loss formulation.

Implementation details. Optimization runs for 500 steps of Adam with learning rate 0.05 on $T = 8$ uniformly sampled frames at 224×224 resolution. The loss weights $(\lambda_A, \lambda_M, \lambda_S, \lambda_D)$ are set via a coarse grid search and then a local Optuna search around the best region on a held-out validation subset. We compare methods with per-clip paired tests (Wilcoxon signed-rank). Hyperparameters, sensitivity sweeps, hardware, and timing are reported in the Appendix.

5.3 Results

Table 1 reports the separation metrics across the three benchmarks. Figure 2 shows an EGTEA Gaze+ clip that a TSM model classifies as *pour water*. The motion mask keeps two of the eight frames (t=3 and t=6), in which the hand operates the faucet above the pot, and discards the frames where the camera shows the whole sink (t=0) or turns away (t=7); each discarded frame is replaced by a copy of its nearest kept frame. The appearance mask appears only on the three frames out of which two frames are kept by M_m and cover the faucet, the pot and the hand. STEP and Extremal Perturbation, which use a single spatio-temporal mask, place their evidence on different regions in nearly every frame, including the frames showing the whole sink and the camera turning away.

Deletion AUC. With the TSM backbone, our method attains the lowest motion deletion AUC on all three datasets and the lowest appearance deletion AUC on EPIC-Kitchens-55 and EGTEA. On EPIC-Kitchens-55, appearance deletion AUC drops to 0.10 versus 0.18 for Extremal Perturbation and 0.23 for STEP, and motion deletion AUC drops to 0.21 versus 0.50 for the Extremal Perturbation temporal extension (EP + M_m). On EGTEA the values are 0.27 (appearance) and 0.25 (motion). On Something-Something V2, motion deletion AUC is again the lowest (0.11 versus 0.39 for EP + M_m), but appearance deletion AUC (0.11) is second to EP + M_m (0.08) and statistically tied with Extremal Perturbation (Table 7). With the TimeSformer backbone, our method has the lowest deletion AUC on every dataset along both axes. Removing the regions and frames selected by our masks therefore lowers target-class confidence more than removing those selected by the baselines in all but one setting, indicating that the masks capture evidence the classifier relies on.

Mask sparsity. On EPIC-Kitchens-55, our appearance mask retains only 10.3% of pixels and our motion mask retains only 24.5% of frames. STEP retains 66.1% of pixels and lacks a temporal mask (using every frame); Unified Mask retains 21.1% of pixels but similarly uses every frame. The pattern recurs on EGTEA (12.4% appearance / 37.9% motion for our method versus 56.4% and all frames for STEP) and on Something-Something V2 (22.8% / 22.1% versus 78.9% and all frames). The ReLU upper-bound area loss therefore performs as designed: when the input does not require its full budget, the optimizer uses less.

Insertion AUC. Our method does not lead on insertion AUC. Baselines with larger masks, such as the STEP variants and Unified Mask, often attain higher insertion scores. On EGTEA, Unified Mask achieves an appearance insertion AUC of 0.81 and STEP + M_m reaches 0.84 on the motion axis, compared to our 0.75 and 0.79. On EPIC-Kitchens-55, STEP + M_m reaches 0.70 motion insertion versus our 0.66. On Something-Something V2, Unified Mask leads appearance insertion (0.39) and STEP + M_m leads motion insertion (0.48), against our 0.31 and 0.25. As formalized in Proposition 4, for binary or near-binary masks insertion AUC tends to reward mask size: a larger mask that contains the evidence keeps confidence high over a longer portion of the sweep. Our appearance masks retain 3.5–$6.4\times$ fewer pixels than STEP and about $2\times$ fewer than Unified Mask, so they are at a disadvantage under this metric, which is why we treat deletion AUC as the primary measure of evidence localization.

Proposition 4 (Insertion AUC favors larger faithful masks when they are forced to be binary i.e important or not)**.** Assuming two masks $M^{(1)}$ and $M^{(2)}$, characterized by active area fractions $0 < a_1 < a_2 < 1$ and strictly satisfy the faithfulness axioms of Definition 1, and assuming the slope of the insertion curve of $M^{(2)}$ over the interval $[a_1, a_2]$ is bounded below by a strictly positive constant $\sigma > 0$. Then the difference between insertion AUC is given by:

$$\mathrm{AUC}_{\mathrm{Ins}}^{(2)} - \mathrm{AUC}_{\mathrm{Ins}}^{(1)} \;\geq\; \tfrac{1}{2}\sigma\,(a_2 - a_1)^2. \tag{12}$$

The accumulated gap remains strictly positive under any circumstance where method 2's supplementary positions contain individually informative predictive signals.

Proposition 5 (Deletion AUC depends on coverage, not size)**.** Suppose $M^{(1)}$ sufficiently covers the model's underlying evidence, where the deletion curve $D^{(1)}$ perfectly converges to the baseline confidence: $D^{(1)}(a^*) = C_{\text{base}}$(fully blurred video). Under these conditions: $\mathrm{AUC}_{\mathrm{Del}}^{(2)} \;=\; \mathrm{AUC}_{\mathrm{Del}}^{(1)}$.

Appendix B.3 formalizes this argument: under perfect-faithfulness assumptions, a larger mask provably enjoys higher insertion AUC by a margin linear in the area gap (Proposition 4), while deletion AUC is invariant to mask size at fixed precision (Proposition 5). Our method outperforms or

matches all baselines on deletion AUC across all three datasets and along both axes, which we view as the more direct evidence of explanation quality supported by the above propositions.

Explanation fidelity (F^+ and F^-). We evaluate faithfulness via evidence retention (F^+) and complementary fidelity (F^-). With the TSM backbone, our method attains the highest motion F^- on all three datasets: 0.82 on EGTEA (versus 0.11 for EP + M_m and 0.12 for STEP + M_m), 0.62 on EPIC-Kitchens-55 (versus 0.20 and 0.21), and 0.62 on Something-Something V2, where STEP + M_m is close (0.61). Removing the frames selected by M_m therefore lowers confidence in the target (verb or class) far more than removing the frames selected by the baselines' temporal masks, indicating that M_m captures the temporal evidence the classifier relies on. Although our appearance masks retain 3.5–$6.4\times$ fewer pixels than STEP, they keep $F^+ \geq 0.97$ on EPIC-Kitchens-55 and EGTEA. The exception is motion F^+ on Something-Something V2 (0.13): keeping only the $\sim 22\%$ of frames selected by M_m is not sufficient to preserve the prediction there, consistent with the lower motion insertion AUC on that dataset.

Ablations. Isolating our factorization, the unfactored Unified Mask (TSM) achieves high appearance insertion AUC (0.81 EGTEA, 0.59 EPIC) and F^+ (0.99, 0.99) simply by retaining all frames and larger spatial footprints (20.2%, 21.1%). Lacking temporal gating, it fails at precise localization, yielding worse appearance deletion AUCs than our factored approach (0.43 vs. 0.27 EGTEA; 0.22 vs. 0.1 EPIC) and offering no motion explanation. This indicates that the dual-mask factorization, rather than the shared perturbation pipeline, drives the deletion gains. Appendix A.3 reports a cumulative component ablation, an ablation of the noun/verb semantic split (App. B.4), the flow-alignment ablation (App. B.6), and hyperparameter sweeps (App. B.11).

Auditing example. Grouping EGTEA predictions by verb and ranking classes by few active M_m frames and high confidence surfaces nominally temporal verbs whose explanations collapse to almost no frames: for *wash*, M_m keeps on average 0.8 of 8 frames while the classifier remains $\approx 90\%$ confident (App. B.9). This flags *wash* as a candidate static-appearance shortcut, a hypothesis that a practitioner can then test by intervention.

Cost. Optimization takes 11.7 s per clip on TSM, comparable to STEP (11.4 s) and EP (10.3 s). Because mask parameters are independent across clips, batching eight clips reduces the amortized cost by about $6.5\times$, to roughly 2 s per clip (App. B.12).

6 Conclusions and Limitations

We frame video action explanation as a disentangled two-mask problem, isolating spatial appearance and temporal dynamics. Our core mechanisms: factored perturbation, STE-based frame replacement, separation loss, and a one-sided ReLU area constraint are individually motivated and jointly necessary to guarantee explanatory precision. Against STEP, Extremal Perturbation, and unified-mask baselines, our method consistently attains the lowest deletion AUC and highest complementary fidelity (F^-) on both axes, while producing masks roughly $5\times$ sparser than the strongest competitor. By separately highlighting exactly *where* and *when* a model looks, our factored explanations provide practitioners with a precise post-hoc auditing tool to expose spurious shortcuts. Our proposed framework serves strictly as a post-hoc auditing tool for frozen video classifiers, producing spatial and temporal masks that expose the model's internal reliance. Because it provides diagnostic evidence rather than inherent interpretability or accuracy improvements, practitioners should use it to surface candidate failure modes and shortcuts, avoiding the misuse of confident-looking saliency maps to arbitrarily validate incorrect predictions. However, the framework remains limited by its reliance on a flow alignment loss for single-head models, which introduces an inductive bias toward motion-rich regions. The factorization separates the two perturbation axes by construction, but it does not force the classifier's representations to separate; measured cross-axis gradients are substantial. Finally, factoring strictly by appearance and motion is inherently restrictive, making the method unsuitable for actions defined by audio cues, gaze patterns, off-frame interactions or long-range interactions that last more than 8 frames.

Acknowledgments and Disclosure of Funding

This work was funded by Oregon State University and its School of Electrical Engineering and Computer Science. We are grateful to Fuxin Li for valuable advice, and to Devika Venkatesh and Dmitriy Shmidt for their help with computational resources.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

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

[3] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian Conference on Image Analysis*, 2003.

[4] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[5] R. Fong, M. Patrick, and A. Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[6] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14(7):1217–1225, 2019.

[7] 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.

[8] T. Hartley, K. Sidorov, C. Willis, and D. Marshall. 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.

[9] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

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

[11] Patrick Knab, Sascha Marton, Philipp J Schubert, Drago Guggiana, and Christian Bartelt. Concepts in motion: Temporal concept bottleneck model for interpretable video classification. *arXiv preprint arXiv:2509.20899*, 2025.

[12] M. Kowal, A. Dave, R. Ambrus, A. Gaidon, K. G. Derpanis, and P. Tokmakov. Understanding video transformers via universal concept discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[13] J. Lee, W. Lee, G.-M. Park, S. T. Kim, and J. Choi. Disentangled concepts speak louder than words: Explainable video action recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[14] Y. Li, M. Liu, and J. M. Rehg. 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)*, 2018.

[15] 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.

[16] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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

[18] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[19] A. Saha, S. Gupta, S. K. Ankireddy, K. Chahine, and J. Ghosh. Exploring explainability in video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[20] 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.

[21] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, and R. Poppe. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

[22] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[23] 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.

[24] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the ACM International Conference on Multimedia*, 2019.

A Appendix

A.1 Reproducibility statement

Code, the hyperparameter search ranges and selected values (App. B.10), the evaluation clip lists, and the TimeSformer checkpoints trained on EPIC-Kitchens-55 and EGTEA Gaze+ are available at `https://github.com/vbhavank/wherewhen`. Frozen SSv2 weights are taken from the official releases of the cited backbones. All results are supported by paired significance tests (App. B.7). Like the classifiers it explains, the method could be used to refine action recognition systems deployed for surveillance; we release it only as an analysis tool for existing models.

A.2 Qualitative Results

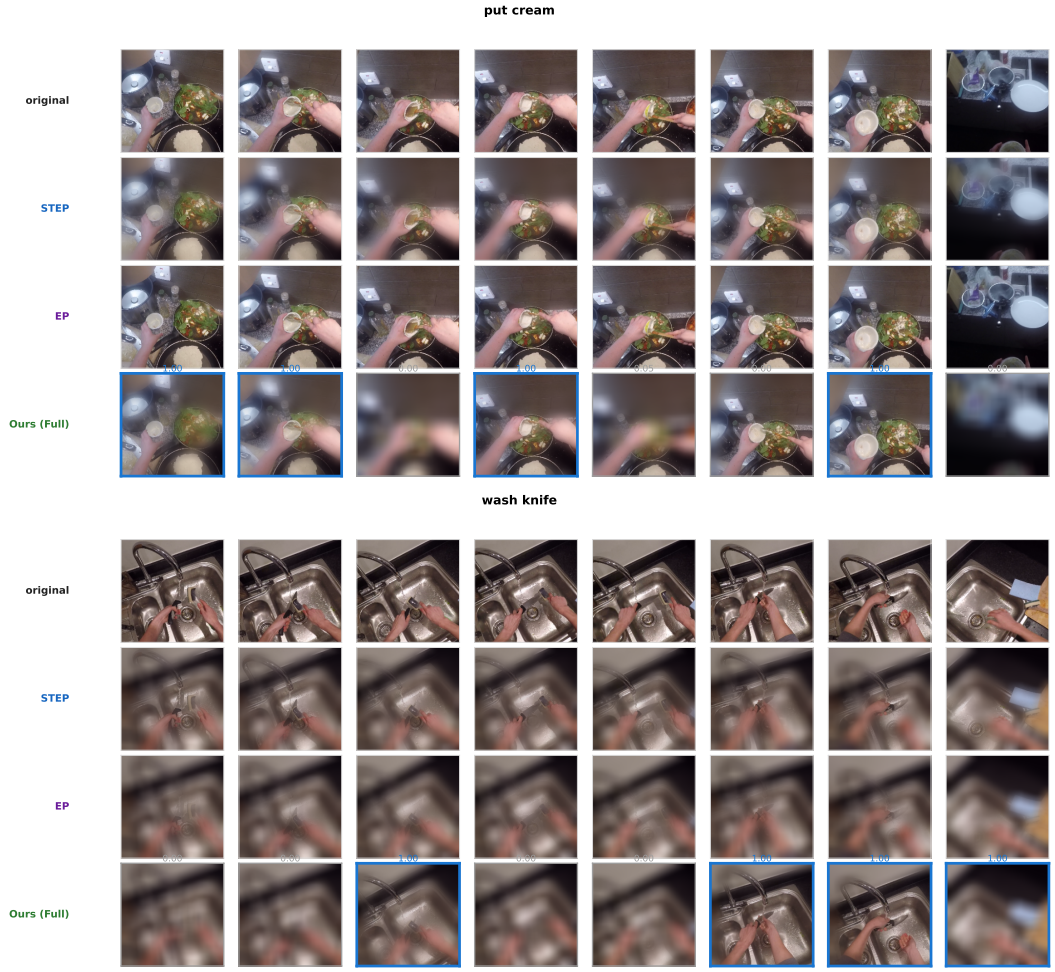


Figure 3: Qualitative comparison of explanation methods on a single Epic-Kitchens clip for action *put-cream* and *wash-knife*. Each column is one of the 8 uniformly sampled frames. **Row 1 (original):** input frames. **Row 2 (STEP [15]):** spatial mask M_a learned by STEP; the highlighted region is shown sharp while the complement is replaced by a Gaussian blur (31×31, $\sigma=20$), matching the perturbation seen by the model. **Row 3 (Extremal Perturbation [5]):** same rendering for the EP spatial mask. **Row 4 (Ours):** our disentangled spatial mask under the same blur-blend rendering; frame borders are blue when the temporal mask $M_m>0.5$ (model-relevant frame) and gray when frozen (nearest-active-frame replacement applied). Values above each frame give the soft M_m weight.

A.3 Ablations

We cumulatively ablate our framework from the STEP baseline on EGTEA (frozen TSM; Figure 4). First, replacing MSE with the ReLU upper-bound area loss prevents non-discriminative masking

by allowing sub-budget use, improving Appearance Insertion AUC (0.53→0.61) and Deletion AUC (0.56→0.52). Next, adding separation loss ($\mathcal{L}_{\text{sep}}$) prevents spatial masking on frozen frames, driving massive faithfulness gains: Appearance Deletion AUC drops to 0.21, Motion Deletion drops drastically from 0.48 to 0.08, and Appearance Insertion increases to 0.96. Finally, the Full Model (incorporating STE hard frame selection) solidifies temporal perturbations, achieving our best Appearance Deletion (0.16) and peak Motion Insertion (0.60), while Motion Deletion remains highly competitive (0.10) and Appearance Insertion normalizes to 0.85.

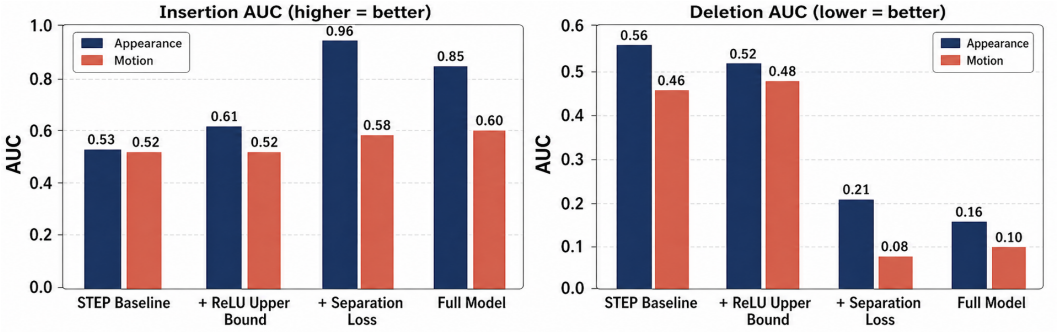


Figure 4: Cumulative ablation on a frozen TSM trained on EGTEA Gaze+. Each bar adds a component on top of the previous configuration.

B Theoretical analysis

This appendix formalizes four claims from the main text. Section B.1 bounds the bias of the Straight-Through Estimator used for hard frame selection (§3.1). Section B.2 characterizes the stationary points of the symmetric MSE area loss and the proposed ReLU upper-bound area loss (§3.2), formalizing why the latter permits sub-budget masks. Section B.3 formalizes the structural advantage of deletion AUC over insertion AUC for sparse explanations (§5).

We adopt the following notation. Vectors and tensors are bold; scalar entries are non-bold and indexed by subscripts. The classifier f is twice continuously differentiable as a function of its input, which holds for all architectures considered in the main text. The clean clip is $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times 3}$, the spatially-perturbed clip is $\mathbf{x}^{\text{app}}$ as in §3.1, and the final perturbed clip is $\hat{\mathbf{x}}$ as in Eq. (6). We write $\nabla_\theta \mathcal{L}$ for the gradient of a scalar loss with respect to a parameter tensor θ, and $\text{sg}(\cdot)$ for the stop-gradient operator (zero in the backward pass, identity in the forward pass).

B.1 Bias of the Straight-Through Estimator for hard frame selection

The frame replacement scheme in §3.1 employs the gradient estimator

$$\tilde{w}_{i\to j} \;=\; \mathbb{1}[j = j^*(i)] \;-\; \text{sg}(w_{i\to j}) \;+\; w_{i\to j}, \tag{13}$$

where $j^*(i) = \arg\max_j w_{i\to j}$ and $w_{i\to j}$ is the soft replacement weight defined in §3.1. The forward pass of Eq. (13) evaluates to the one-hot vector $\mathbf{e}_{j^*(i)}$, while the backward pass yields the gradient of the soft weights:

$$\frac{\partial \tilde{w}_{i\to j}}{\partial m_k} \;=\; \frac{\partial w_{i\to j}}{\partial m_k}.$$

For any downstream scalar loss $\mathcal{L}$ dependent on the perturbed clip, the STE backward computation produces

$$\widehat{\nabla_{m_k}\mathcal{L}} \;=\; \sum_{i,j} \frac{\partial \mathcal{L}}{\partial \tilde{w}_{i\to j}} \cdot \frac{\partial w_{i\to j}}{\partial m_k}, \tag{14}$$

whereas the gradient of the true forward function (one-hot $\arg\max$) is zero almost everywhere and undefined on a measure-zero set of ties. Following [9], the relevant comparison is therefore not against the "true" gradient, which provides no signal, but against the gradient of a surrogate: the loss evaluated on the fully soft mixture.

Setup. Let $\hat{\mathbf{x}}_{\text{soft}}(\mathbf{m})$ denote the perturbed clip computed using soft weights w instead of the STE-corrected $\tilde{w}$, and let $\mathcal{L}_{\text{soft}}(\mathbf{m}) := \mathcal{L}(\hat{\mathbf{x}}_{\text{soft}}(\mathbf{m}))$. The STE gradient (14) equals $\nabla_{\mathbf{m}}\mathcal{L}_{\text{soft}}$ evaluated through $\hat{\mathbf{x}}_{\text{soft}}$, but in our pipeline the loss is evaluated at $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\text{hard}}$, the discretized clip. The bias is formalized as

$$\text{Bias}(\mathbf{m}) \;=\; \nabla_{\mathbf{m}}\mathcal{L}_{\text{soft}}\big|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\text{hard}}} \;-\; \nabla_{\mathbf{m}}\mathcal{L}_{\text{soft}}\big|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\text{soft}}}. \tag{15}$$

Let L be a Lipschitz constant of $\nabla\mathcal{L}_{\text{soft}}$ with respect to $\hat{\mathbf{x}}$ under the ℓ_1 norm, viewed as a function from clips to gradients with respect to $\mathbf{m}$. Such a constant exists for any twice continuously differentiable classifier on a bounded input domain.

Intuitively, this bounds the deviation of the STE gradient from that of a strictly soft pipeline. The bias is minimized when the soft argmax distribution is highly peaked (confident) or when the candidate replacement frames are visually similar.

Proposition 1 (STE bias bound)**.** For each frame index i, let $\Delta_i := w_{i\to j^*(i)} - w_{i\to j^{(2)}(i)}$ denote the gap between the largest and second-largest replacement weight, and let $\delta_i := \max_{j\neq j^*(i)} \|\mathbf{x}^{\text{app}}_{j^*(i)} - \mathbf{x}^{\text{app}}_{j}\|$ denote the largest pairwise distance between candidate replacement frames. Then

$$\big\|\text{Bias}(\mathbf{m})\big\| \;\le\; L \cdot \sum_{i\,:\,M_{m}{}_i^{\text{STE}}=0} (1-\Delta_i)\cdot\delta_i. \tag{16}$$

Proof. Fix a frame i marked frozen by the temporal STE gate ($M_{m}{}_i^{\text{STE}} = 0$, replacement active). The hard and soft replacement frames at position i are

$$\hat{\mathbf{x}}_i^{\text{hard}} = \mathbf{x}^{\text{app}}_{j^*(i)}, \qquad \hat{\mathbf{x}}_i^{\text{soft}} = \sum_j w_{i\to j}\,\mathbf{x}^{\text{app}}_j.$$

We need to measure the visual distance between the "hard" frame (if we forced the model to pick exactly one frame) and the "soft" blended frame (the weighted average the math thinks we chose).

First, we subtract the soft frame from the hard frame. Because the soft frame is just a mixture of the winning frame and all the losing frames, the winning frame perfectly cancels out its own portion of the blend. We are left with just a sum of all the "losing" frames ($j \neq j^*$), weighted by their probabilities ($w_{i\to j}$), multiplied by how visually different they are from the winner:

$$\|\hat{\mathbf{x}}_i^{\text{hard}} - \hat{\mathbf{x}}_i^{\text{soft}}\| \;=\; \Big\| \sum_{j\neq j^*(i)} w_{i\to j}\big(\mathbf{x}^{\text{app}}_{j^*(i)} - \mathbf{x}^{\text{app}}_{j}\big)\Big\|$$

Calculating the exact visual difference for every single losing frame is too complex, so we establish an upper bound. The Triangle Inequality allows us to pull the absolute value (norm) inside the summation. Instead of tracking exact differences, we assume the worst-case scenario. Let δ_i represent the absolute biggest visual difference between the winning frame and *any* candidate frame. By replacing the exact distances with this maximum distance, we guarantee our formula acts as a strict ceiling ($\le$):

$$\Big\| \sum_{j\neq j^*(i)} w_{i\to j}\big(\mathbf{x}^{\text{app}}_{j^*(i)} - \mathbf{x}^{\text{app}}_{j}\big)\Big\| \;\le\; \sum_{j\neq j^*(i)} w_{i\to j}\cdot\delta_i$$

Because the weights (w) are probabilities, the sum of the probabilities for every candidate frame must equal 1 (100%). Therefore, the combined probability of all the *losing* frames is simply 1 minus the probability of the *winning* frame ($w_{i\to j^*(i)}$). This neatly collapses the entire summation:

$$\sum_{j\neq j^*(i)} w_{i\to j}\cdot\delta_i \;=\; \big(1 - w_{i\to j^*(i)}\big)\cdot\delta_i$$

We define the gap (Δ_i) as the exact difference in probability between the first-place winner and the second-place runner-up ($j^{(2)}$):

$$w_{i\to j^*(i)} \;=\; w_{i\to j^{(2)}(i)} + \Delta_i$$

Because a probability can never be negative ($w_{i\to j^{(2)}(i)} \ge 0$), the winning probability must be *at least* as large as the gap itself: $w_{i\to j^*(i)} \ge \Delta_i$.

If the winning probability is greater than or equal to Δ_i, then subtracting it from 1 flips the inequality sign, giving us:

$$1 - w_{i \to j^*(i)} \;\leq\; 1 - \Delta_i$$

By substituting this gap logic back into our collapsed equation from, we arrive at the final pointwise bound. The visual error between the hard choice and the soft choice is mathematically guaranteed to be less than or equal to the visual difference of the worst frame (δ_i) multiplied by the uncertainty of the decision ($1 - \Delta_i$):

$$\|\hat{\mathbf{x}}_i^{\text{hard}} - \hat{\mathbf{x}}_i^{\text{soft}}\| \;\leq\; (1 - \Delta_i) \cdot \delta_i.$$

Summing the norms over all frozen frames bounds the ℓ_1 distance between the hard and soft clips. Applying the ℓ_1-Lipschitz continuity of $\nabla \mathcal{L}_{\text{soft}}$ then yields Eq. (16). □

Analysis of the Bias Bound. The bound comprises two factors per frozen frame, capturing independent properties of the replacement scheme. The factor $(1 - \Delta_i)$ quantifies the entropy of the soft argmax; a confident replacement (i.e., a large margin between the top two weights) minimizes this term. The factor δ_i measures the visual discrepancy among candidate replacement frames; for temporally redundant segments, δ_i is small, rendering the specific replacement choice inconsequential.

These factors interact with the temperature annealing schedule and replacement geometry. Because temperature annealing is applied to the *mask* sigmoid rather than the replacement weights directly, it does not force $\Delta_i \to 1$ asymptotically. Instead, it pushes the mask values toward $\{0, 1\}$, yielding an exactly one-hot $\tilde{w}$ during the forward pass and restricting the active set to a sparse subset of frames. Within this active set, the proximity Gaussian $\exp(-\|i - j\|^2 / 2\sigma_R^2)$ (where $\sigma_R = 2$) separates the nearest active neighbor from the second nearest by a non-trivial multiplicative factor (e.g., $\exp(-1/8)/\exp(-1/2) \approx 1.46$ for distances 1 and 2). Consequently, Δ_i is lower-bounded by a constant determined solely by the active set's geometry. Furthermore, δ_i remains small when the nearest-active-frame replacement selects a visually similar source, which is typical for slowly varying video segments.

B.2 Stationary points of the area constraints

In our method, we want the neural network to highlight important parts of a video. We give it a "budget" (a target size for the highlight mask, called v). We want to prove mathematically why using a standard Mean Squared Error (MSE) loss forces the mask to match this budget *exactly*, whereas using a ReLU loss allows the mask to be *smaller* than the budget if it finds the important features early. Let us characterize the stationary points of the joint loss $\mathcal{L}_{\text{pred}}(\mathbf{m}) + \lambda \mathcal{L}_{\text{area}}(\mathbf{m}, v)$ under both the symmetric MSE and the proposed ReLU upper-bound area constraints.

Setup. Let $\mathbf{m} \in [0, 1]^N$ denote a generic mask (representing M_a flattened, or M_m, or any sub-block thereof). Let $\mathcal{L}_{\text{pred}} : [0, 1]^N \to \mathbb{R}_{\geq 0}$ be continuously differentiable with bounded gradients, $\|\nabla \mathcal{L}_{\text{pred}}(\mathbf{m})\| \leq L$ for some constant L. In practice $\mathcal{L}_{\text{pred}}$ has plateau regions where $\nabla \mathcal{L}_{\text{pred}}(\mathbf{m}^*) = \mathbf{0}$ once the mask covers enough of the model's evidence.

Let $v \in (0, 1)$ denote the target area fraction and $t_i \in \{0, 1\}$ the binary target with $t_i = 1$ for $i \geq (1 - v)N$ in sorted order. We write $\mathbf{m}^{\text{sort}}$ for the descending sort of $\mathbf{m}$ and abbreviate $a := \|\mathbf{m}\|_1 / N$ for the realized area fraction.

The symmetric and one-sided area losses are

$$\mathcal{L}_{\text{area}}^{\text{MSE}}(\mathbf{m}, v) \;=\; \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{m}_i^{\text{sort}} - t_i\right)^2, \qquad \mathcal{L}_{\text{area}}^{\text{ReLU}}(\mathbf{m}, v) \;=\; \frac{1}{N} \sum_{i=1}^{N} \left[\text{ReLU}(\mathbf{m}_i^{\text{sort}} - t_i)\right]^2. \quad (17)$$

The symmetric area loss formulation invariably produces masks that saturate the allocated budget. As the penalty weight λ increases, the realized mask area converges to the target budget at a rate of $O(1/\lambda)$.

Proposition 2 (MSE forces budget saturation, with quantitative rate)**.** Let $\mathbf{m}_\lambda^*$ be an interior stationary point of $\mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{area}}^{\text{MSE}}$. Then the realized area satisfies

$$|a_\lambda^* - v| \;\leq\; \frac{L\sqrt{N}}{2\lambda}. \quad (18)$$

In particular, $a_\lambda^* \to v$ as $\lambda \to \infty$, and the gap closes at rate $O(1/\lambda)$ (with constant scaling as $\sqrt{N}$).

Proof. At an interior stationary point, the first-order condition dictates that the total gradient must equal zero:

$$\nabla \mathcal{L}_{\text{pred}}(\mathbf{m}_\lambda^*) + \lambda \nabla \mathcal{L}_{\text{area}}^{\text{MSE}}(\mathbf{m}_\lambda^*, v) = \mathbf{0}.$$

Rearranging this equivalence yields $\lambda \nabla \mathcal{L}_{\text{area}}^{\text{MSE}}(\mathbf{m}_\lambda^*, v) = -\nabla \mathcal{L}_{\text{pred}}(\mathbf{m}_\lambda^*)$. By taking the norm of both sides and applying the assumed gradient bound $\|\nabla \mathcal{L}_{\text{pred}}\| \leq L$, we deduce that the norm of the area loss gradient must be strictly bounded:

$$\|\nabla \mathcal{L}_{\text{area}}^{\text{MSE}}(\mathbf{m}_\lambda^*, v)\| \leq \frac{L}{\lambda}.$$

This mathematically enforces that as the penalty weight λ grows, the area loss gradient is driven toward zero. To evaluate this gradient norm, we compute the partial derivative of the MSE area loss with respect to a specific mask parameter m_k. Applying the chain rule to the quadratic penalty yields $\partial \mathcal{L}_{\text{area}}^{\text{MSE}} / \partial m_k = (2/N)(\mathbf{m}_{i_k}^{\text{sort}} - t_{i_k})$, where i_k represents the rank index of position k after the descending sort.

Computing the squared Euclidean norm (ℓ_2 norm) of this gradient vector entails summing the squares of its components over all N dimensions:

$$\|\nabla \mathcal{L}_{\text{area}}^{\text{MSE}}\|_2^2 \;=\; \frac{4}{N^2} \sum_{i=1}^{N} (\mathbf{m}_i^{\text{sort}} - t_i)^2 \;\leq\; \frac{L^2}{\lambda^2}.$$

Equating this algebraic expansion to our previously established upper bound $(L/\lambda)^2$ and isolating the summation yields a bound on the sum of squared differences: $\sum_i (\mathbf{m}_i^{\text{sort}} - t_i)^2 \leq N^2 L^2 / (4\lambda^2)$.

We now relate this squared Euclidean distance to the absolute difference in total area. By invoking the Cauchy–Schwarz inequality, which asserts that the sum of elements (ℓ_1 norm) is bounded by the square root of the sum of their squares (ℓ_2 norm) scaled by the dimension N, we obtain:

$$\left| \sum_i (\mathbf{m}_i^{\text{sort}} - t_i) \right| \;\leq\; \sqrt{N \cdot \sum_i (\mathbf{m}_i^{\text{sort}} - t_i)^2} \;\leq\; \frac{NL\sqrt{N}}{2\lambda}.$$

To conclude the proof, we recognize that the realized area a_λ^* is simply the mean of the sorted mask $(1/N) \sum_i \mathbf{m}_i^{\text{sort}}$, and the target budget v is the mean of the binary targets $(1/N) \sum_i t_i$.

$$a_\lambda^* = \frac{1}{N} \sum_{i=1}^{N} \mathbf{m}_i^{\text{sort}}$$

$$v = \frac{1}{N} \sum_{i=1}^{N} t_i$$

We want to find the absolute difference between these two quantities. Substituting our definitions gives:

$$|a_\lambda^* - v| \;=\; \left| \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{m}_i^{\text{sort}} \right) - \left(\frac{1}{N} \sum_{i=1}^{N} t_i \right) \right|$$

Because $\frac{1}{N}$ is a positive constant shared by both terms, we can factor it completely outside of the absolute value, combining the summations into a single term:

$$|a_\lambda^* - v| \;=\; \frac{1}{N} \left| \sum_{i=1}^{N} (\mathbf{m}_i^{\text{sort}} - t_i) \right|$$

From our earlier application of the Cauchy-Schwarz inequality, we established a strict upper bound for this absolute sum: $\left| \sum_{i=1}^{N} (\mathbf{m}_i^{\text{sort}} - t_i) \right| \leq \frac{NL\sqrt{N}}{2\lambda}$. We substitute this upper bound into our equation:

$$|a_\lambda^* - v| \;\leq\; \frac{1}{N} \cdot \left(\frac{NL\sqrt{N}}{2\lambda} \right)$$

The N in the denominator perfectly cancels the N in the numerator of the bound, leaving us withEq. (18):

$$|a^*_\lambda - v| \;\le\; \frac{L\sqrt{N}}{2\lambda}$$

□

Conversely, the one-sided area loss permits the mask to converge below the target budget, provided the prediction loss is minimized. Once the mask area strictly satisfies the budget constraint, the area penalty yields a zero gradient, thereby allowing the realized sparsity to be dictated entirely by the prediction loss landscape.

Proposition 3 (ReLU permits sub-budget stationary points independent of λ)**.** Suppose there exists $\mathbf{m}^* \in [0,1]^N$ with $\nabla\mathcal{L}_{\text{pred}}(\mathbf{m}^*) = \mathbf{0}$ and $\mathbf{m}_i^{\text{sort},*} \le t_i$ for all i (so $a^* \le v$). Then $\mathbf{m}^*$ is a stationary point of $\mathcal{L}_{\text{pred}} + \lambda\mathcal{L}_{\text{area}}^{\text{ReLU}}$ for every $\lambda > 0$.

Proof. We analyze the stationary condition by first computing the partial derivative of the ReLU area loss. Applying the chain rule to the squared ReLU function gives:

$$\frac{\partial\mathcal{L}_{\text{area}}^{\text{ReLU}}}{\partial m_k} \;=\; \frac{2}{N}\,\text{ReLU}\big(\mathbf{m}_{i_k}^{\text{sort}} - t_{i_k}\big).$$

The critical property of the ReLU function is that it evaluates to zero for any non-positive argument. Under the hypothesis that the proposed mask strictly satisfies the budget constraint (i.e., $\mathbf{m}_i^{\text{sort},*} \le t_i$ for all i), the argument $(\mathbf{m}_i^{\text{sort},*} - t_{i_k})$ is inherently non-positive across all dimensions. Consequently, the ReLU argument returns zero, and the gradient evaluates to exactly zero everywhere.

Substituting this zero-gradient result back into the global first-order stationary condition yields:

$$\nabla\mathcal{L}_{\text{pred}}(\mathbf{m}^*) + \lambda \cdot \mathbf{0} = \mathbf{0}.$$

Because the area penalty gradient vanishes completely, the optimality of the point relies solely on the prediction loss landscape. Since we posited by hypothesis that $\nabla\mathcal{L}_{\text{pred}}(\mathbf{m}^*) = \mathbf{0}$, this equality holds true independent of the magnitude of the hyperparameter λ. □

What the two propositions say jointly. Proposition 2 formalizes the empirical observation that MSE-area-loss masks always realize close to the budget. The symmetric penalty penalizes the mask for being smaller than v as well as for being larger, so any interior stationary point is pushed toward $a = v$ at rate $O(1/\lambda)$. Proposition 3 shows that the ReLU form decouples realized area from budget: once the mask is sub-budget, the area term contributes nothing to the gradient, and the realized sparsity is determined entirely by the geometry of $\mathcal{L}_{\text{pred}}$. This is the behavior we exploit in Table 1, where realized appearance-mask sparsity falls well below the budget set during optimization.

B.3 Insertion AUC has a structural advantage for larger masks

We now formalize the analytical claim introduced in Section 5: for explanation masks whose values are driven to 0 or 1, the Insertion AUC metric implicitly rewards mask size, whereas the Deletion AUC metric strictly resists this confounding factor. The effect is specific to such (near-)binary masks and does not arise in the same form for continuous attribution maps such as saliency maps. Perturbation methods that optimize masks toward binary values (our method, Extremal Perturbation and STEP, all of which anneal or threshold their masks toward $\{0,1\}$) fall in the binary regime analyzed here.

Setup. Fix an input clip and a designated target classification class. For a given explanatory mask M evaluated over the $N := THW$ spatial-temporal positions, the standard insertion procedure sequentially reveals the top-$\lfloor \alpha N \rfloor$ positions in descending order of mask magnitude (utilizing consistent permutation for tie-breaking). All remaining unrevealed positions are subjected to a baseline blur, and the target-class probability is subsequently evaluated. We denote this resulting evaluation curve as $C^{(M)} : [0,1] \to [0,1]$ and the normalized area-fraction of the active mask as $a(M) := |\{i : M_i = 1\}|/N$. By definition, the boundary conditions evaluate to $C^{(M)}(0) = C_{\text{base}}$ (the fully blurred baseline prediction) and $C^{(M)}(1) = C_{\text{full}}$ (the original unperturbed prediction), both of which are strictly independent of the chosen mask M.

We comparatively analyze two distinct masks, $M^{(1)}$ and $M^{(2)}$, characterized by active area fractions $0 < a_1 < a_2 < 1$. We stipulate that both masks share an underlying *position importance ordering* π. This implies their active sets are strictly nested: method 1 isolates the top $a_1 N$ positions, while method 2 isolates the top $a_2 N$ positions, both following the exact π-order. Furthermore, tie-breaking among all inactive positions ($M_i = 0$) relies on the identical permutation for both methods. This specific configuration mathematically models the scenario where method 2's supplementary active positions are precisely the marginal features that method 1 evaluated as minimally important and subsequently excluded.

Definition 1 formalizes the intuitive premise that a mask captures the entirety of the model's predictive evidence and nothing extraneous. Condition (iii) represents the core substantive constraint: spatial-temporal positions excluded from the mask yield strictly zero informational value to the classifier until the complete input tensor is restored.

Definition 1 (Faithfulness)**.** A mask M with corresponding insertion curve $C^{(M)}$ is defined as faithful if it satisfies the following axioms:

(i) $C^{(M)}$ is strictly non-decreasing across the domain $[0, 1]$ (revealing additional visual information never deteriorates the prediction confidence);
(ii) $C^{(M)}$ is concave over the domain $[0, a(M)]$ (each newly revealed active position contributes a weakly lesser increase in confidence than the preceding position);
(iii) $C^{(M)}(\alpha) = C^{(M)}(a(M))$ for all $\alpha \in [a(M), 1)$, with the original confidence $C^{(M)}(1) = C_{\text{full}}$ recovered exclusively as a discontinuous jump at $\alpha = 1$.

Condition (iii) is the primary engine of the faithfulness assumption: positions deemed inactive by the mask contain zero predictive signal until the full context is forcibly restored. This constraint is mathematically stronger than the generic inequalities occasionally utilized in statements of this metric, allowing for rigorous geometric bounds in the ensuing proofs. While empirical explanation methods rarely satisfy Condition (iii) perfectly, we address the practical implications of this relaxation below.

Notation. Let us define the baseline-adjusted confidence curves as $S(\alpha) := C^{(1)}(\alpha) - C_{\text{base}}$ and $T(\alpha) := C^{(2)}(\alpha) - C_{\text{base}}$. Both functions are non-negative and non-decreasing, anchored at $S(0) = T(0) = 0$ and terminating at $S(1) = T(1) = C_{\text{full}} - C_{\text{base}}$. For the fractional domain $\alpha \leq a_1$, the insertion procedures of both methods reveal mathematically identical spatial-temporal coordinates (because their importance orderings are shared, and method 1's active set acts as a strict prefix to method 2's). Consequently, $S(\alpha) = T(\alpha)$ precisely on $[0, a_1]$. We denote the pointwise deviation between the curves as $\Delta(\alpha) := T(\alpha) - S(\alpha) \geq 0$.

Proposition 4 (Insertion AUC favors larger faithful masks)**.** Assuming the setup defined above wherein both $M^{(1)}$ and $M^{(2)}$ strictly satisfy the faithfulness axioms of Definition 1, and assuming the slope of the insertion curve $C^{(2)}$ over the interval $[a_1, a_2]$ is bounded below by a strictly positive constant $\sigma > 0$, then:

$$\text{AUC}_{\text{Ins}}^{(2)} - \text{AUC}_{\text{Ins}}^{(1)} \;\geq\; \tfrac{1}{2}\sigma\,(a_2 - a_1)^2. \tag{19}$$

The accumulated gap remains strictly positive under any circumstance where method 2's supplementary positions contain individually informative predictive signals.

Proof. We begin by computing the explicit difference between the accumulated areas under the respective insertion curves. By definition, the Insertion AUC for method k is the integral of its raw curve over the fractional domain $\alpha \in [0, 1]$. Separating the baseline offset yields:

$$\text{AUC}_{\text{Ins}}^{(k)} = \int_0^1 C^{(k)}(\alpha)\, d\alpha = C_{\text{base}} + \int_0^1 (C^{(k)}(\alpha) - C_{\text{base}})\, d\alpha.$$

Thus, the difference in AUC between the two methods reduces purely to the integral of the pointwise gap $\Delta(\alpha) = T(\alpha) - S(\alpha)$:

$$\text{AUC}_{\text{Ins}}^{(2)} - \text{AUC}_{\text{Ins}}^{(1)} = \int_0^1 \Delta(\alpha)\, d\alpha.$$

To isolate the region of divergence, we partition the domain of integration into three distinct intervals based on the active mask fractions a_1 and a_2:

$$\int_0^1 \Delta(\alpha)\, d\alpha = \underbrace{\int_0^{a_1} \Delta(\alpha)\, d\alpha}_{\text{Interval 1}} + \underbrace{\int_{a_1}^{a_2} \Delta(\alpha)\, d\alpha}_{\text{Interval 2}} + \underbrace{\int_{a_2}^{1} \Delta(\alpha)\, d\alpha}_{\text{Interval 3}}.$$

In Interval 1 ($[0, a_1]$), both methods sequentially reveal identical spatial-temporal features. Thus, $S(\alpha) = T(\alpha)$, rendering $\Delta(\alpha) = 0$, and ensuring the first integral evaluates exactly to zero. In Interval 3 ($[a_2, 1]$), the pointwise gap $\Delta(\alpha)$ is guaranteed to be non-negative because $T(\alpha) \geq S(\alpha)$ pointwise by the definition of the nested active sets. Therefore, the third integral contributes a value strictly greater than or equal to zero.

Because the first integral is zero and the third is non-negative, we can establish a strict mathematical lower bound by analyzing Interval 2 exclusively:

$$\text{AUC}_{\text{Ins}}^{(2)} - \text{AUC}_{\text{Ins}}^{(1)} \;\geq\; \int_{a_1}^{a_2} \Delta(\alpha)\, d\alpha.$$

Within the fractional domain $(a_1, a_2]$, method 2 continues to reveal genuinely informative positions, driving its curve upward. Conversely, method 1 has entirely exhausted its active mask subset. By invoking Faithfulness Condition (iii) on $M^{(1)}$, we establish that any newly revealed positions past a_1 contribute exactly zero change to the model's confidence until $\alpha = 1$. Consequently, $S(\alpha)$ plateaus entirely: $S(\alpha) = S(a_1)$ for all $\alpha \in (a_1, 1)$. The gap function within Interval 2 thereby simplifies cleanly to $\Delta(\alpha) = T(\alpha) - S(a_1)$.

We are provided the hypothesis that the slope of $C^{(2)}$ (and equivalently T) over the interval $[a_1, a_2]$ is bounded below by the constant $\sigma > 0$. Because the curves intersect at the boundary such that $T(a_1) = S(a_1)$, we can project a linear lower bound onto the trajectory of $T(\alpha)$: $T(\alpha) - S(a_1) \geq \sigma(\alpha - a_1)$.

Substituting this linear geometric bound back into the integral yields:

$$\int_{a_1}^{a_2} \Delta(\alpha)\, d\alpha \;\geq\; \int_{a_1}^{a_2} \sigma(\alpha - a_1)\, d\alpha.$$

In other words, because the small mask flatlines and the big mask climbs at slope σ, they form a triangle of "free area" on the graph. The width of this triangle is $(a_2 - a_1)$. The height is $\sigma(a_2 - a_1)$. The area of a triangle is $\frac{1}{2} \cdot \text{base} \cdot \text{height}$. Evaluating this standard polynomial integral computes to $\frac{1}{2}\sigma(\alpha - a_1)^2$ evaluated from boundary a_1 to a_2, directly resulting in $\frac{1}{2}\sigma(a_2 - a_1)^2$. This derivation finalizes the formal quadratic bound presented in Eq. (19). The established quadratic bound is structurally conservative because it entirely omits the accumulated area from the third interval $[a_2, 1]$. □

Consequently, the Insertion AUC metric exhibits an inherent structural bias that strictly favors spatially larger masks, even when controlling for per-position precision. Because a larger active mask possesses a wider fractional domain over which to incrementally increase the model's confidence, it mathematically accumulates a definitively greater area under the evaluation curve. In stark contrast, the Deletion AUC metric intrinsically penalizes an artificial inflation of mask size. Once a proposed mask successfully encompasses the entirety of the model's predictive evidence, further expansion of the active mask cannot mathematically depress the deletion curve any further. Consequently, Deletion AUC functions strictly as a measure of evidence coverage, remaining invariant to arbitrary increases in absolute mask size.

Proposition 5 (Deletion AUC depends on coverage, not size)**.** Assume $M^{(1)}$ and $M^{(2)}$ represent faithful masks (per Definition 1) that share the position importance ordering, subject to $0 < a_1 < a_2 < 1$. Suppose $M^{(1)}$ sufficiently covers the model's underlying evidence, formally defined such that there exists a fractional point $a^* \in (0, a_1]$ where the deletion curve $D^{(1)}$ perfectly converges to the baseline confidence: $D^{(1)}(a^*) = C_{\text{base}}$. Under these conditions:

$$\text{AUC}_{\text{Del}}^{(2)} \;=\; \text{AUC}_{\text{Del}}^{(1)}. \tag{20}$$

Proof. The deletion evaluation procedure at a given fractional threshold α dictates that the top-$\lfloor \alpha N \rfloor$ positions of the mask, ranked strictly in descending order, are removed and replaced with a baseline blur. The target-class probability is then recalculated.

Because masks $M^{(1)}$ and $M^{(2)}$ perfectly share the underlying position importance ordering π, both methods inherently prioritize the exact same spatial-temporal features for initial deletion. Therefore, for any fraction α within the interval $[0, a_1]$, the deletion operations drop identical sets of pixels. This strictly guarantees that their respective deletion curves are pointwise equivalent throughout this initial domain: $D^{(1)}(\alpha) = D^{(2)}(\alpha)$ for all $\alpha \in [0, a_1]$.

Invoking the coverage hypothesis, the curve $D^{(1)}$ is known to reach the absolute baseline confidence C_{base} at some fractional point $a^* \leq a_1$. By fundamental definition, a deletion curve is monotonically non-increasing and is strictly bounded from below by C_{base}. Because $D^{(1)}$ intersects its absolute minimum at a^*, it is geometrically constrained to remain perfectly horizontal thereafter: $D^{(1)}(\alpha) = C_{\text{base}}$ for all subsequent fractions $\alpha \in [a^*, 1)$.

Crucially, this exact geometric plateau must uniformly apply to $D^{(2)}$ over the interval $[a^*, 1)$. Because $D^{(1)}$ and $D^{(2)}$ mandate identical pixel removals up to a_1, and given that $a^* \leq a_1$, the curve $D^{(2)}$ has similarly been driven down to the baseline C_{base} precisely at $\alpha = a^*$. Further obfuscation of features beyond a^* cannot rationally reduce the model's confidence below the baseline probability. Thus, $D^{(2)}(\alpha) = C_{\text{base}}$ identically for all $\alpha \in [a^*, 1)$.

For the remaining fractional domain $\alpha \in [a_2, 1]$, both explanatory methods have completely exhausted their proposed active positions. They proceed to sequentially delete inactive (zero-weighted) positions governed entirely by the shared tie-breaking permutation. As theoretically established, both functions merely trace the C_{base} minimum floor throughout this entire tail process until $\alpha = 1$.

Because $D^{(1)}(\alpha) = D^{(2)}(\alpha)$ over the domain $[0, a^*]$, and both curves strictly equate to C_{base} over $[a^*, 1]$, we conclude that $D^{(1)} \equiv D^{(2)}$ globally across the entire closed interval $[0, 1]$. Integrating identical functional curves inevitably yields identical areas, cementing Eq. (20). □

The asymmetry between insertion and deletion. Once an explanatory mask achieves total coverage of the model's operative evidence, expanding its boundaries changes the deletion curve exclusively at thresholds $\alpha < a^*$, a domain where the curves of any structurally covering masks natively agree by construction. Deletion AUC thus rigorously evaluates whether a mask captures the prediction support, deliberately ignoring extraneous inclusions. The contrapositive statement proves equally diagnostic: if two masks produce substantially divergent deletion AUCs, we can mathematically conclude that at least one candidate fundamentally fails to capture the core evidence.

The structural advantage of Insertion AUC for larger masks fundamentally arises from the directional placement of the mask's "tail" (the large quantity of low-importance background pixels). We characterize this phenomenon as *asymmetric early saturation*. Every explanation mask can be conceptually partitioned into a high-importance "head" (the core visual evidence) and a low-importance "tail" (the surrounding background context). The mathematical asymmetry between the two metrics lies entirely in how they integrate this tail.

The Insertion Phenomenon (Multiplying by High Confidence). Under the insertion protocol, the mask reveals its head first. Once the critical evidence is revealed, the model's target-class confidence rapidly saturates near its maximum ($C \approx 1$). For a large, dense mask, the remaining evaluation steps consist of revealing its large tail. Because the confidence is already saturated, this tail is integrated over a high-confidence plateau. Geometrically, the tail creates a wide integration window that is multiplied by a high altitude, racking up strictly positive area simply because the mask takes longer to finish revealing pixels.

The Deletion Phenomenon (Multiplying by Zero). Conversely, the deletion protocol hides the head first. Once the critical evidence is destroyed, the model is effectively blinded to the target class, and confidence crashes near its baseline minimum ($C \approx 0$). When the large mask proceeds to hide its massive tail, it operates within a "dead zone." Integrating over this zero-confidence plateau contributes negligibly to the total area. Ultimately, Deletion AUC strictly isolates the efficiency of the "head" because the tail is mathematically multiplied by zero once the evidence is destroyed. Insertion AUC, by building the evidence first, evaluates the tail at maximum confidence. This geometrically

guarantees that larger masks receive artificially inflated scores purely as a function of their absolute pixel count, rather than the precision of their explanation.

Collectively, these two propositions rigorously harden the rhetorical claim presented in Section 5. Insertion AUC mathematically favors enlarged masks even when localized precision is controlled (Proposition 4). Thus, achieving a superior insertion AUC remains highly ambiguous, consistent with either a highly precise compact mask or a loosely imprecise sprawling mask. Deletion AUC actively discriminates between these divergent scenarios (Proposition 5). Our proposed method's consistent dominance in deletion AUC across all three diverse datasets consequently serves as the far more resilient and diagnostic empirical finding.

Note, the presented propositions fundamentally presume *exact* geometric faithfulness as defined in Definition 1, an idealized property that no empirical explanation algorithm flawlessly achieves. The substantive analytical takeaway is qualitative rather than strictly numerical: Insertion AUC inherently harbors a documented geometric bias toward mask size, a vulnerability that Deletion AUC fundamentally resists. Real-world explanatory methods occupy a gradient between perfect faithfulness and arbitrary feature selection; however, the relative divergence between these two metrics on that continuum dependably tracks the limiting theoretical behavior established here.

Empirical check without the idealized faithfulness assumption. Definition 1(iii) is idealized. To test the size bias directly, we build synthetic masks in which size and precision vary independently: at precision p, a fraction p of the active positions is taken from a gradient-saliency ranking and the rest are drawn uniformly at random. We grow each mask from 5% to 60% of all spatio-temporal positions at fixed precision.

Table 2: Insertion vs. deletion AUC under a $12\times$ increase in mask size at fixed precision.

Precision	Ins. (5%)	Ins. (60%)	Change	Del. (5%)	Del. (60%)	Change
0.10	0.282	0.328	$+16.4\%$	0.268	0.236	-11.8%
0.30	0.293	0.396	$+35.5\%$	0.258	0.228	-11.7%
0.50	0.303	0.444	$+46.6\%$	0.252	0.224	-11.1%
0.70	0.311	0.484	$+55.5\%$	0.244	0.223	-8.8%
0.90	0.314	0.505	$+60.8\%$	0.242	0.218	-9.9%

At every precision level, size moves insertion AUC 1.4–$6.3\times$ more than deletion AUC, and the gap widens as precision increases, i.e., exactly for the compact, high-precision masks we aim to produce. This supports the qualitative conclusion of Propositions 4–5 without relying on Definition 1(iii).

B.4 Ablation of the noun/verb semantic split

We keep every architectural component and replace the separate noun and verb objectives with a single blended confidence objective, so neither mask receives an axis-specific semantic signal (50 examples from EGTEA data on the TSM architecture).

Table 3: Effect of removing the semantic split. Δ = (w/o split) $-$ (full).

Metric	Full	w/o semantic split	Δ
App. Del. AUC $\downarrow$	0.128	0.152	$+0.024$
Mot. Del. AUC $\downarrow$	0.234	0.285	$+0.051$
App. F^- $\uparrow$	0.323	0.295	-0.028
Mot. F^- $\uparrow$	0.656	0.616	-0.040
App. Ins. AUC $\uparrow$	0.630	0.673	$+0.043$
Mot. Ins. AUC $\uparrow$	0.703	0.778	$+0.075$
M_m size (%)	≈ 27	36.3	$+9.3$

Removing the split worsens deletion AUC and F^- on both axes. Insertion AUC rises, but together with a $\approx 34\%$ larger temporal mask, consistent with the size bias of insertion (App. B.3). Semantic supervision therefore contributes materially to the separation; the perturbation architecture alone does not determine it.

B.5 Empirical cross-axis gradients

We measure, at convergence, the gradient magnitude of each loss with respect to each mask on 200 clips using a TSM model trained on the EGTEA dataset. The leak ratio is computed per clip as $\|\partial\mathcal{L}_{\text{verb}}/\partial M_a\|/\|\partial\mathcal{L}_{\text{noun}}/\partial M_a\|$ for M_a and $\|\partial\mathcal{L}_{\text{noun}}/\partial M_m\|/\|\partial\mathcal{L}_{\text{verb}}/\partial M_m\|$ for M_m; 0 would mean no cross-axis influence.

Table 4: Gradient magnitudes at convergence. Ratios are computed per clip and then aggregated, so they are not the ratio of the column means.

Mask	Intended gradient	Cross-axis gradient	Leak ratio (mean / median)
M_a	$\|\partial\mathcal{L}_{\text{noun}}/\partial M_a\| = 1.45\times10^{-2}$	$\|\partial\mathcal{L}_{\text{verb}}/\partial M_a\| = 1.38\times10^{-2}$	0.88 / 0.85
M_m	$\|\partial\mathcal{L}_{\text{verb}}/\partial M_m\| = 6.50\times10^{-1}$	$\|\partial\mathcal{L}_{\text{noun}}/\partial M_m\| = 3.20\times10^{-1}$	5.31 / 1.23

Cross-axis gradients are of the same order as the intended ones. For M_m the mean is inflated by clips whose intended verb gradient is nearly zero at convergence (median 1.07×10^{-3}), so the median is more representative. These measurements do not support a claim that gradients separate by semantic axis cleanly. Future work aims to improve the formulation with the use of stop-gradients to prevent gradient leakage.

B.6 How much does optical flow drive the SSv2 motion mask?

Table 5: Flow-alignment ablation on 50 SSv2 clips (TSM). r: Pearson correlation between M_m and per-frame flow magnitude.

Metric	With flow (λ_F=1.312)	Without flow (λ_F=0)
Mot. Ins. AUC ↑	**0.334**	0.186
Mot. Del. AUC ↓	0.152	**0.096**
Mot. F^+ ↑	**0.465**	0.175
Mot. F^- ↑	0.613	**0.647**
M_m size (%)	67.1	37.7
$r(M_m, \text{flow})$	0.37	-0.212

Flow alignment biases M_m toward broader, more sufficient temporal explanations (higher insertion and F^+) at the cost of sparsity and deletion selectivity. The moderate correlation ($r = 0.37$) shows that M_m is influenced by, but does not simply copy, the flow prior. A model-derived replacement for the flow prior is left to future work.

B.7 Seed variance and paired significance tests

Each method is optimized with three seeds. For paired comparisons we compute, per clip, the difference between our method and each baseline, oriented so that positive values favor our method (lower is better for deletion, higher for insertion). We report the mean difference and the Wilcoxon signed-rank p-value.

Table 6: Mean ± std over three seeds (TSM) on the SSV2 dataset.

Dataset	Method	App. Ins. ↑	Mot. Ins. ↑	App. Del. ↓	Mot. Del. ↓
SSv2	EP	0.321 ± 0.002	-	0.143 ± 0.001	-
	EP + M_m	0.266 ± 0.003	0.402 ± 0.007	0.085 ± 0.003	0.399 ± 0.007
	STEP	0.347 ± 0.003	-	0.172 ± 0.001	-
	STEP + M_m	0.357 ± 0.009	0.488 ± 0.001	0.147 ± 0.005	0.444 ± 0.012
	Unified Mask	0.409 ± 0.008	-	0.186 ± 0.005	-
	Ours	0.318 ± 0.007	0.25 ± 0.003	0.11 ± 0.004	0.112 ± 0.002

B.8 Input and mask resolution

Nothing in the framework fixes the input or mask resolution. At 512×512, peak memory grows by $\approx 4.1\times$ and runtime by 6.7–7.6× relative to 224×224. Explanation quality at 512×512 cannot be assessed meaningfully with our classifiers as they were trained at 224×224 and their confidence on

Table 7: Paired comparison of Ours against each baseline (TSM). Each cell gives the mean per-item difference, oriented so that positive values favor our method, and the Wilcoxon signed-rank p-value; *: $p < 0.05$. Items are clips for SSv2 (n=200) and EGTEA (n=400), and videos for EPIC-55 (n=89–93; per-segment scores averaged within each video and over seeds). Motion metrics are omitted for baselines without a motion mask, as in Table 1. Differences are computed on the seed-averaged per-item scores used for this test, so they can differ slightly from differences of the rounded means in Table 1.

Dataset	Baseline	App. Ins.	Mot. Ins.	App. Del.	Mot. Del.
SSv2	EP + M_m	$+0.052^*$ $p=2.3\times10^{-3}$	-0.146^* $p=3.6\times10^{-9}$	-0.028^* $p=3.6\times10^{-5}$	$+0.286^*$ $p=2.1\times10^{-28}$
	EP	-0.020 $p=0.30$	–	$+0.002$ $p=0.51$	–
	STEP + M_m	-0.039^* $p=8.6\times10^{-3}$	-0.232^* $p=6.7\times10^{-19}$	$+0.034^*$ $p=0.033$	$+0.331^*$ $p=8.0\times10^{-32}$
	STEP	-0.029^* $p=0.021$	–	$+0.059^*$ $p=1.9\times10^{-4}$	–
	Unified Mask	-0.091^* $p=4.4\times10^{-6}$	–	$+0.072^*$ $p=5.2\times10^{-15}$	–
EGTEA	EP + M_m	$+0.110^*$ $p=3\times10^{-22}$	$+0.106^*$ $p=7\times10^{-15}$	$+0.060^*$ $p=6\times10^{-13}$	$+0.368^*$ $p=3\times10^{-51}$
	EP	$+0.089^*$ $p=1\times10^{-16}$	–	$+0.075^*$ $p=7\times10^{-18}$	–
	STEP + M_m	-0.010 $p=0.41$	-0.028^* $p=0.03$	$+0.115^*$ $p=7\times10^{-43}$	$+0.449^*$ $p=8\times10^{-59}$
	STEP	$+0.006$ $p=0.06$	–	$+0.142^*$ $p=3\times10^{-41}$	–
	Unified Mask	-0.075^* $p=1\times10^{-5}$	–	$+0.107^*$ $p=2\times10^{-42}$	–
EPIC-55	EP + M_m	$+0.015^*$ $p=0.034$	$+0.100^*$ $p=8.4\times10^{-14}$	$+0.073^*$ $p=2.5\times10^{-15}$	$+0.319^*$ $p=1.4\times10^{-15}$
	EP	$+0.029^*$ $p=1.1\times10^{-4}$	–	$+0.084^*$ $p=1.1\times10^{-15}$	–
	STEP + M_m	-0.152^* $p=1.4\times10^{-14}$	-0.061^* $p=3.3\times10^{-10}$	$+0.122^*$ $p=2.9\times10^{-16}$	$+0.427^*$ $p=2.8\times10^{-16}$
	STEP	-0.138^* $p=1.5\times10^{-12}$	–	$+0.134^*$ $p=8.1\times10^{-16}$	–
	Unified Mask	-0.151^* $p=5.2\times10^{-16}$	–	$+0.127^*$ $p=1.5\times10^{-16}$	–

the *unperturbed* clip drops from 0.855 to 0.215. Even in this setting, raising the seed-mask resolution from 32 to 64 improves motion insertion AUC from 0.307 to 0.526 and appearance insertion AUC from 0.142 to 0.160 on the EGTEA dataset with the TSM model, suggesting that finer seed masks help localize small objects in egocentric video when paired with a classifier trained at the target resolution.

B.9 Auditing example on EGTEA Gaze+

We group 200 unfiltered EGTEA test clips by predicted verb and rank classes by few active M_m frames and high verb confidence. A nominally temporal class whose motion explanation repeatedly collapses to almost no frames while confidence stays high is a candidate for a static-appearance shortcut.

For *wash*, the motion mask keeps fewer than one frame on average while the classifier is $\approx 90\%$ confident, suggesting the verb is recognized largely from static cues (e.g., sink, running water) rather than the washing motion. The mask alone does not establish a shortcut. It produces a testable

Table 8: Verb classes ranked by few active motion frames (of 8) and high confidence (TSM).

Predicted verb	Clips	Mean active M_m frames	Mean verb confidence
Wash	5	0.80	0.895
Cut	10	1.40	0.849
Move around	4	1.50	0.776
Close	3	1.67	0.875

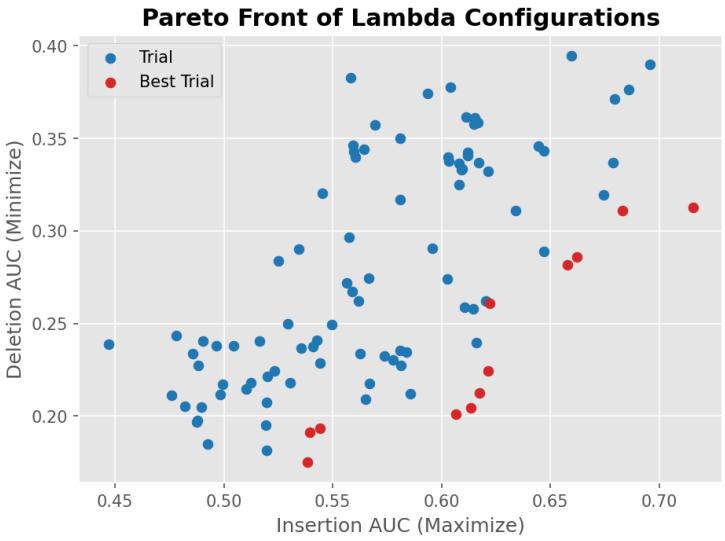


Figure 5: Figure showing how hyperparameters $(\lambda_A, \lambda_M, \lambda_S, \lambda_D)$ were picked with a held out validation set using a combination of grid search and Bayesian optimization on Optuna for minimizing deletion AUC and maximizing Insertion AUC across 100 trials on the EGTEA Gaze+ dataset with the TSM model.

hypothesis: freezing all frames to the single most appearance-salient frame should leave *wash* confidence largely intact, whereas a motion-dependent class should drop. Conversely, preserving appearance while altering the trajectory tests whether the model depends on motion at all.

B.10 Hyperparameters

Figure 5 shows how hyperparameters $(\lambda_A, \lambda_M, \lambda_S, \lambda_D)$ were picked for the EPIC-Kitchens 55 dataset using a held out validation set, across 100 trials. Table 9 shows the hyperparameters that resulted from the Bayesian optimization on a held-out set.

B.11 Hyperparameter sensitivity

We vary one hyperparameter at a time around our tuned configuration, holding all others fixed, on 20 clips per dataset with three optimization seeds per setting (Table 10). Seed-to-seed standard deviations are small relative to the trends (at most 0.035 on EGTEA and 0.062 on SSv2), so the effects below are not noise.

Area budgets. The appearance budget v_a has almost no effect: across a $3\times$ range all four AUCs vary by at most 0.006, consistent with the upper-bound area loss being inactive when the optimized mask is already smaller than the budget. The motion budget v_m leaves insertion unchanged but trades away deletion selectivity as it grows (App. Del. $0.162 \rightarrow 0.224$), and motion Fidelity$^-$ drops at the largest budget ($0.71 \rightarrow 0.59$).

Separation weight λ_D. λ_D is the most influential parameter and controls a trade-off between sufficiency and selectivity on both datasets: increasing it raises insertion AUC (SSv2 Mot. Ins. $0.042 \rightarrow 0.431$) and motion Fidelity$^+$ ($0.04 \rightarrow 0.64$), while deletion AUC also increases. Very small λ_D yields near-zero deletion AUC on SSv2, but only because the motion mask collapses: motion

Table 9: Optimization hyperparameters for our disentangled mask (M_a, M_m). Loss weights $(\lambda_A, \lambda_M, \lambda_S, \lambda_D, \lambda_F)$ were tuned on a held-out split via Bayesian optimization. Optimizer settings, schedule, and temperature anneal are identical across datasets and backbones.

(a) Loss weights and target areas (per dataset)			
	EPIC-Kitchens-55	EGTEA Gaze+	Something-Something V2
λ_A (M_a area)	13.85	0.3841	0.133
λ_M (M_m area)	9.44	0.786	0.263
λ_S (L_K smoothness)	0.143	0.0107	0.0877
λ_D (disentanglement)	7.026	38.88	38.88
λ_F (flow alignment)	N/A	N/A	1.312
ν_a (target spatial area for M_a)	0.20	0.53	0.6
ν_m (target temporal area for M_m)	0.50	0.61	0.6
(b) Optimizer and schedule (shared across datasets)			
Optimizer	Adam, $(\beta_1, \beta_2) = (0.9, 0.999)$, ϵ=10^{-8}		
Trainable parameters	$\boldsymbol{\ell}_a \in \mathbb{R}^{T\times 32\times 32}$, $\boldsymbol{\ell}_m \in \mathbb{R}^T$ (logits)		
Learning rate	$\alpha = 0.05$ (constant)		
Weight decay	0		
Iterations	N=500 per clip		
Logit initialization	$\boldsymbol{\ell}_a, \boldsymbol{\ell}_m \leftarrow -1.0$ ($\sigma(-1)\approx$0.27)		
Mask upsample	32×32 → 224×224 via Gaussian (factor 7×)		
L_K kernel	spatial 31×31 (stride 31), temporal 3 (stride 1)		
(c) Temperature annealing schedule			
Mask parameterization	$M = \sigma(\tau_i \cdot \boldsymbol{\ell})$		
Annealing form	linear in iteration		
τ_i	$\tau_i = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \frac{i}{N-1}$		
$\tau_{\min} \to \tau_{\max}$	1.0 → 10.0 over 500 iterations		
Final-mask temperature	$\tau_{\max} = 10.0$ (binarizing)		

insertion and Fidelity$^+$ fall to almost zero. A low deletion AUC on its own is therefore not evidence of a good explanation, and our tuned value lies in the regime where both sufficiency and selectivity are reasonable.

Flow weight λ_F. On SSv2, across a $50\times$ range of λ_F App. Ins. stays within 0.25–0.33 and Mot. Ins. within 0.30–0.42; only the weakest flow prior (λ_F=0.1) noticeably lowers motion insertion, and values between 1 and 2 form a plateau. The flow term is used only for single-head classifiers, so it has no effect on EGTEA (identical results at every value).

Optimization steps. Insertion AUC increases monotonically with the number of steps, with diminishing returns beyond our setting of 500 on SSv2 (App. Ins. +0.011 from 500 to 750). Deletion AUC rises with more steps as the masks become more sufficient, and 100 steps are clearly too few (SSv2 appearance Fidelity$^+$ 0.55 vs. 0.97 at 500).

B.12 Compute and timing

All optimization-based explanation methods were evaluated using a single NVIDIA DGX. Because our framework and the baselines act as post-hoc explainers applied to a frozen classifier, no model retraining is required. The computational cost per example scales primarily with the number of optimization iterations ($N_{\text{iter}} = 500$) and the complexity of the forward and backward passes through the chosen video backbone.

On the EGTEA Gaze+ and EPIC-Kitchens-55 datasets using the TSM backbone ($T = 8$ frames), our full factored method requires an average of 11.7 seconds per video clip. The computational overhead of our nearest-active-frame temporal replacement strategy adds only minor latency compared to standard single-mask baselines: STEP requires 11.4s (12.8s with the M_m temporal extension), and Extremal Perturbation requires 10.3s (10.4s with temporal extension). The Unified Mask ablation, which optimizes only a single spatiotemporal mask without the appearance-motion factorization, bypasses the temporal gating entirely and is consequently the fastest at 7.3s per example.

Table 10: Hyperparameter sensitivity (TSM; 20 clips $\times$ 3 seeds, mean AUC). One hyperparameter is varied at a time; † marks the tuned value used for all other rows. This study uses its own clip subset, so absolute values are not directly comparable to Table 1.

Dataset	Param.	Value	App. Ins. ↑	Mot. Ins. ↑	App. Del. ↓	Mot. Del. ↓
EGTEA	v_a	0.26	.699	.730	.193	.270
		0.40	.696	.729	.196	.267
		0.53†	.700	.730	.192	.270
		0.66	.694	.729	.197	.270
		0.79	.696	.731	.198	.270
	v_m	0.31	.695	.727	.162	.267
		0.46	.693	.727	.179	.270
		0.61†	.700	.730	.192	.270
		0.76	.698	.732	.212	.277
		0.92	.700	.734	.224	.283
	λ_D	19.4	.604	.686	.122	.244
		29.2	.658	.710	.159	.256
		38.9†	.700	.730	.192	.270
		48.6	.710	.771	.214	.284
		58.3	.724	.787	.231	.295
	Steps	250	.639	.649	.153	.212
		375	.682	.706	.177	.238
		500†	.700	.730	.192	.270
		625	.706	.750	.196	.292
		750	.731	.769	.207	.309
SSv2	λ_D	1	.105	.042	.018	.026
		5	.128	.042	.027	.032
		20	.265	.246	.088	.128
		38.9†	.281	.395	.134	.209
		50	.297	.431	.133	.230
	λ_F	0.1	.327	.297	.077	.133
		1	.284	.397	.134	.193
		1.31†	.281	.395	.134	.209
		2	.305	.421	.139	.222
		5	.253	.373	.133	.168
	Steps	100	.115	.016	.031	.010
		250	.226	.195	.107	.090
		500†	.281	.395	.134	.209
		750	.292	.448	.143	.237

For Something-Something V2, which inherently demands more complex spatial-temporal processing, our full method requires 18.1 seconds per example. Given the evaluation subset sizes ($N = 400$ for EGTEA, $N = 9385$ for EPIC-Kitchens-55, and $N = 200$ for SSv2), generating the complete suite of explanations for a single method on a single dataset takes roughly 1 to 2 hours for EGTEA and 32hours for EPIC-Kitchens. Factoring in all baseline evaluations, multiple backbones (TSM and TSFormer), and hyperparameter grid searches, the total compute budget for the experiments reported in this paper was approximately 400 to 500 GPU hours.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction accurately state the paper's main contributions: a post-hoc framework for disentangling appearance and motion explanations in video action recognition, a factored perturbation design with spatial and temporal masks, a separation evaluation protocol, and empirical evaluation across EPIC-Kitchens-55, EGTEA Gaze+, and Something-Something V2. The claims are supported by the method, formal analysis, and results table. The scope is mostly clear: the strongest claims apply to frozen video classifiers and to the evaluated datasets/models, while the single-head setting is handled through a flow-alignment heuristic rather than a full semantic verb/noun decomposition.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The paper discusses several limitations, including the reliance on optical-flow alignment for single-head datasets, the fact that this is a heuristic rather than a structurally guaranteed semantic decomposition, and the possibility that appearance/motion factorization may not capture actions driven by audio, gaze, off-frame context, or longer temporal dependencies than the sampled clip. The paper also reports computational cost and per-example timing, which addresses efficiency.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: The paper includes several theoretical results, including propositions on STE pointwise discrepancy, MSE budget saturation, and ReLU sub-budget stationary points. The main text states the results and references appendix sections for proofs. The assumptions are stated near the corresponding propositions, and the results are numbered and cross-referenced.
4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: The paper provides the main information needed to reproduce the experimental results: datasets, model families, baselines, mask parameterization, perturbation procedure, optimization length, optimizer, learning rate, frame count, input resolution, evaluation metrics, and hyperparameter-selection procedure. It also states that code, hyperparameters, held-out grids, and checkpoints will be released upon acceptance.
5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: We provide open access to the code. The datasets used in our experiments are public benchmarks, and we release the code and trained checkpoints.
6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: We specify the main experimental details needed to understand the results, including the datasets, evaluated model architectures, baselines, evaluation metrics, input resolution, number of sampled frames, optimizer, learning rate, number of optimization steps, and the use of Bayesian optimization on a held-out validation subset to select loss weights. We also report subset sizes in Table 1.
7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: We report aggregate metrics over sampled evaluation subsets, All results are supported by paired significance tests.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We report the compute resources used for the experiments, including the GPU setup, approximate per-example runtime, and total compute budget. The appendix states that experiments were run on a single NVIDIA DGX, with approximately 11.7 seconds per EGTEA/EPIC clip using TSM and 18.1 seconds per SSv2 example, and an estimated total compute budget of roughly 400–500 GPU hours.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe the research conforms to it. Our work develops a post-hoc explanation method for existing video action recognition models and evaluates it on established public datasets. We do not introduce new human-subject data, collect private information, or deploy the system in a real-world decision-making setting. We also preserve submission anonymity and acknowledge limitations and potential risks associated with over-interpreting post-hoc explanations.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss several positive societal impacts, including shortcut auditing, robustness forecasting, clinical video analysis, egocentric assistive technology, dataset diagnostics, and comparison with human attention. However, the current paper does not sufficiently discuss potential negative societal impacts. In particular, we should more explicitly address risks such as over-trusting post-hoc explanations, using explanations to justify unreliable video classifiers, privacy concerns in egocentric or clinical video, and possible misuse in surveillance or high-stakes decision-making.

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We do not release a high-risk model, scraped dataset, generative model, or pretrained foundation model that would require special controlled-release safeguards. Our work is a post-hoc explanation method evaluated on established public video benchmarks, and we do not introduce new data collection or release a dataset with new privacy or safety risks. We plan to release code and experiment artifacts for reproducibility, but these do not substantially increase misuse risk beyond the existing public datasets and classifiers.

12. **Licenses for existing assets**

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

    Answer: [No]

    Justification: We cite the original papers for the datasets, models, and baselines used, including EPIC-Kitchens-55, EGTEA Gaze+, Something-Something V2, TSM, TimeSFormer, STEP, and Extremal Perturbation. However, the current paper does not explicitly list the licenses, versions, URLs, or terms of use for these assets.

13. **New assets**

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

Answer: [Yes]

Justification: We release new code and experiment artifacts, including trained checkpoints. We do not introduce a new dataset.

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

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

Answer: [N/A]

Justification: We do not conduct crowdsourcing experiments or new research with human subjects. The work uses existing public video action-recognition datasets and does not collect new participant data, provide new instructions to human annotators, or compensate workers as part of this study.

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

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

Answer: [N/A]

Justification: We do not conduct crowdsourcing experiments or new research with human subjects. The work uses existing public video action-recognition datasets and does not collect new participant data or interact with study participants. Therefore, IRB approval or equivalent review is not applicable to this study.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: Our core method development does not use LLMs as an important, original, or non-standard component. The proposed approach is based on post-hoc perturbation optimization, factored appearance and motion masks, separation losses, and flow alignment.