BEYOND CORRESPONDENCE: UNDERSTANDING RESPONSE VERIFICATION IN IMAGE–EVENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Image–event retrieval matches static images with the brightness changes recorded by event cameras. Candidate verification can refine learned similarity rankings. Understanding its effects requires examining the response cues that distinguish candidates and their influence on initially correct matches. We present a controlled evaluation framework that holds trained models and their top-five candidates fixed while comparing image-derived responses with observed event responses. We compare acquisition-informed and static response representations and test whether combining each response score with the initial similarity score improves on response-only ranking. Experiments with three models on two controlled benchmarks show that acquisition-informed verification improves mean bidirectional top-1 accuracy over initial rankings. Preserving coarse temporal structure aids verification on N-Caltech101, while static gradients outperform acquisition-informed responses for the EventBind and CEIA retrieval models on mini N-ImageNet. With a validation-selected weight held fixed, score fusion improves three-run mean bidirectional top-1 accuracy by 1.60–6.03 percentage points over response-only ranking across eight test settings, while also outperforming initial rankings. A CEIA case study shows how fusion reduces displacement of initially correct matches. These findings provide an empirical basis for evaluating response representations and their integration with learned correspondence in image–event retrieval. Code is available in an anonymous repository.

1 INTRODUCTION

We study *paired image–event instance retrieval*: each image and its corresponding event recording form a one-to-one positive pair, and a query from either modality retrieves its paired observation from the other. Event cameras asynchronously report per-pixel brightness changes, with each event carrying a location, timestamp, and polarity indicating whether brightness increased or decreased (Gallego et al., 2022). Retrieval therefore matches image appearance with the change-driven measurements in its paired recording.

Existing image–event alignment models map the two modalities into a common representation space. Building on Contrastive Language–Image Pretraining (CLIP) (Radford et al., 2021), EventBind (Zhou et al., 2024) and CLIP-based Event–Image Alignment (CEIA) (Xu et al., 2025) rank candidate pairs by learned feature similarity, which we call the *correspondence score*. Assigning the highest score to the unique correct pair remains challenging. Published retrieval evaluations show a gap between top-1 accuracy and top-5 recall (Zhou et al., 2024; Xu et al., 2025). For paired-instance retrieval, this gap identifies correct matches ranked second to fifth, motivating a verification stage to refine their order. Candidate verification has long complemented initial rankings in image retrieval (Philbin et al., 2007; Cao et al., 2020).

Physical and structural priors offer concrete ways to construct this second comparison. Event-formation models relate brightness changes to image gradients and motion (Gallego et al., 2022). Using available acquisition information, we predict brightness-change activity from a candidate image and organize the observed events into a compatible representation. We call these representations *responses*; comparing them in a separate *response space* gives an *acquisition-informed verification*

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

score. Spatial activity maps and local temporal summaries provide further cues for constructing verification representations (Gehrig et al., 2019; Sironi et al., 2018).

Verification can correct ranking errors and displace initially correct matches. We therefore investigate three connected questions about verification *after an initial learned ranking*: whether it improves retrieval, which physical or structural cues yield useful response representations, and whether combining response similarity with the initial correspondence score improves retrieval over response-only reranking.

We use a *correspondence–response* interface for controlled evaluation: a trained model selects the top five candidates, and fixed image and event mappings construct responses for comparison. The model and candidate membership remain fixed across comparisons. We compare response representations under the same ranking rule and, with a given response held fixed, test whether combining its score with learned correspondence improves on response-only reranking. We evaluate N-Caltech101 (Orchard et al., 2015) and mini N-ImageNet (Kim et al., 2021), which pair displayed source images with event recordings captured under controlled camera motion. Alongside EventBind and CEIA, we use Point-Pattern Unbalanced Optimal Transport (PP-UOT), a local-correspondence model trained from scratch, as a controlled initial ranker. Our primary metric is bidirectional top-1 accuracy, the average of the image-to-event and event-to-image results.

We conduct spatial controls on both benchmarks and acquisition-order controls on N-Caltech101. On mini N-ImageNet, where each recording's position within the motion cycle is unspecified, we compare acquisition-informed verification with a *static differential response* based on spatial gradients of the image and a map of accumulated event counts, using shared spatial preprocessing. Additional controlled-motion experiments examine the retrieval value of acquisition-informed scores after static evidence is available.

Finally, we evaluate *score fusion*, a weighted combination of correspondence and response scores. We select the weight using PP-UOT validation data and hold it fixed in follow-up test comparisons. CEIA on mini N-ImageNet illustrates its role: acquisition-informed response-only ranking introduces more errors than it corrects in image-to-event retrieval. Combining the same response with correspondence substantially reduces the loss of initially correct matches and raises accuracy above the initial ranking.

Our main contributions are:

- **We establish that response verification can improve paired retrieval after learned correspondence ranking.** Across the three models and two benchmarks, acquisition-informed reranking improves three-run mean bidirectional top-1 accuracy by 0.64–18.24 percentage points over the initial ranking. Mean changes are positive in all six settings, with 95% confidence intervals excluding zero in five.
- **We show that static spatial responses can outperform acquisition-informed verification on mini N-ImageNet, while temporal organization is useful on N-Caltech101.** On mini N-ImageNet, static verification exceeds acquisition-informed verification by 3.57 and 4.61 percentage points on average for EventBind and CEIA, respectively, with higher bidirectional test accuracy in all three evaluated training runs of each model. On N-Caltech101, experiments that merge or reorder responses from different portions of acquisition show the retrieval value of their coarse temporal organization.
- **We show that combining verification with learned correspondence improves test retrieval beyond either score alone.** With a validation-selected weight held fixed, fusion improves three-run mean bidirectional top-1 accuracy by 1.60–6.03 percentage points over response-only ranking. The eight tested settings cover acquisition-informed responses across all three models and both benchmarks, plus static responses for EventBind and CEIA on mini N-ImageNet. Fusion exceeds both the initial and corresponding response-only rankings in every evaluated training run and retrieval direction.

2 Related Work

Cross-modal and event representations. Contrastive learning provides comparable image and multimodal representations (Radford et al., 2021; Girdhar et al., 2023). Event representations en-

code spatial activity and temporal context (Sironi et al., 2018; Gehrig et al., 2019), while paired pretraining transfers image knowledge to event encoders (Yang et al., 2023). EventCLIP adapts CLIP to event recognition through event-frame representations (Wu et al., 2023). EventBind and CEIA align image and event embeddings for cross-modal comparison (Zhou et al., 2024; Xu et al., 2025).

Candidate verification and physical consistency. Image retrieval combines initial rankings with local spatial verification (Philbin et al., 2007; Noh et al., 2017; Cao et al., 2020). Ji et al. (2023) apply candidate reranking to event-to-image place recognition using a learned pairwise similarity classifier. Event-formation models support simulation and motion-conditioned comparison (Rebecq et al., 2018; Gallego et al., 2018a;b). We study paired-instance retrieval with the trained ranker and candidate membership held fixed, using deterministic response comparisons. We evaluate verification utility, compare acquisition-informed and static responses, and test whether combining response and correspondence scores improves on response-only ranking. Appendix A expands these connections.

3 METHOD

3.1 CORRESPONDENCE–RESPONSE RETRIEVAL

We pair learned correspondence with deterministic response-space comparison (Fig. 1). The resulting scores support response-only reranking and fusion reranking within a fixed shortlist.

Initial ranking. For a query q, let $\mathcal{G}$ contain M opposite-modality observations. The initial model ranks them by decreasing correspondence score $S_{\mathrm{corr}}(q, g)$:

$$\pi_{\mathrm{corr}}(q) = (g_1, \ldots, g_M) = \mathrm{rank}_{g \in \mathcal{G}}\, S_{\mathrm{corr}}(q, g). \tag{1}$$

The first K observations form the candidate set $\mathcal{C}_K(q) = \{g_1, \ldots, g_K\}$. Each checkpoint supplies its own candidate set for each query. The checkpoint and candidate membership remain fixed across comparisons.

Response-space comparison. For an image–event pair (I, E), modality-specific mappings construct compatible responses:

$$R_I = \mathcal{T}_I(I), \qquad R_E = \mathcal{T}_E(E), \qquad S_{\mathrm{resp}}(I, E) = s(R_I, R_E). \tag{2}$$

The fixed mappings retain spatial activity, directional gradients, or ordered acquisition groups using shared spatial and channel conventions. Section 3.2 defines their construction and similarity s.

Response-only reranking. For either retrieval direction, $S_{\mathrm{resp}}(q, g_r)$ denotes the response score of the corresponding image–event pair. Let σ_q sort these scores over the K candidates in descending order. The verified ranking is

$$\pi_{\mathrm{final}}(q) = \big(g_{\sigma_q(1)}, \ldots, g_{\sigma_q(K)}, g_{K+1}, \ldots, g_M\big). \tag{3}$$

Reranking preserves shortlist membership and all lower-ranked positions. The primary protocol uses $K = 5$ as a common candidate budget across models and datasets; Appendix G.5 evaluates other sizes.

Combining correspondence and response similarity. To test whether fusion improves retrieval over response-only reranking, we standardize each score family X over the same K candidates of a query:

$$z_q(X_r) = \frac{X_{q,r} - \mu_q(X)}{\sigma_q(X) + \varepsilon_z}, \tag{4}$$

where $\mu_q(X)$ and $\sigma_q(X)$ are the mean and population standard deviation of those candidate scores, and $\varepsilon_z = 10^{-8}$. The fused score for candidate g_r is

$$S_{\mathrm{fuse}}(q, g_r) = \lambda z_q(S_{\mathrm{corr},r}) + (1 - \lambda) z_q(S_{\mathrm{resp},r}), \qquad \lambda \in [0, 1]. \tag{5}$$

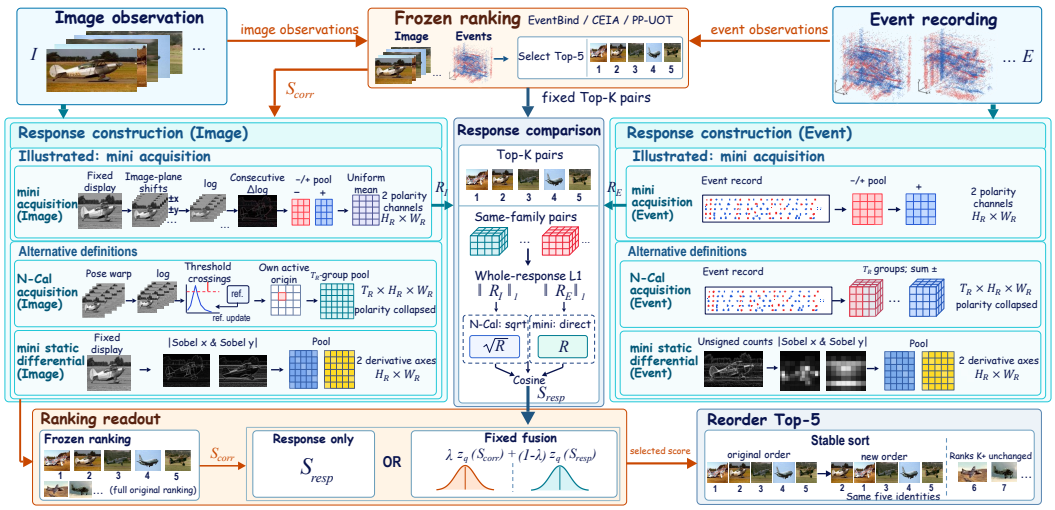


Figure 1: **Correspondence–response retrieval.** A frozen model selects top-K candidates by correspondence similarity S_{corr} ($K = 5$). Compatible image and event responses use direction-averaged spatial–polarity activity on mini N-ImageNet, ordinal acquisition groups on N-Caltech101, or static gradients on mini N-ImageNet. Whole-response ℓ_1 normalization and cosine comparison yield S_{resp}; N-Caltech101 additionally applies an elementwise square root before comparison. The readout uses S_{resp} alone or fuses the two shortlist-standardized scores, with the validation-selected correspondence weight fixed at $\lambda = 0.25$. Stable sorting preserves candidate membership, initial order under exact ties, and ranks $K+1, \ldots, M$. The rank change is schematic; query and gallery roles swap between retrieval directions.

Sorting this score defines σ_q in Eq. 3. The endpoints $\lambda = 0$ and $\lambda = 1$ give response-only and initial correspondence ordering, respectively. For the follow-up test comparisons, we use the correspondence weight $\lambda = 0.25$, selected on PP-UOT validation data and held fixed across the evaluated models and response constructions (Sec. 4.4). Fusion and response-only reranking use the same response scores and candidates; exact ties in these comparisons retain the initial order.

3.2 Response Representations for Verification

Acquisition-informed responses. The fixed acquisition specification ξ comprises the documented motion family, parameters calibrated on training data, and response-construction settings. For a candidate image I, let $W_{\xi,k}(I)$ be its appearance at sensing state k and $L_{\xi,k}(I)$ the corresponding log-intensity representation. We construct

$$R_I^{\xi} = \mathcal{A}_{\xi}(I) = \Psi_{\xi}\Big(\{L_{\xi,k}(I)\}_{k=0}^{K_{\xi}}\Big), \qquad R_E = \mathcal{R}(E), \tag{6}$$

where K_{ξ} is the number of sensing-state transitions. The operator Ψ_{ξ} accumulates brightness-change activity across the selected spatial and acquisition dimensions, and $\mathcal{R}$ aggregates observed events into matching dimensions.

N-Caltech101: rotational response. We maintain a log-intensity reference across the documented rotational sequence (Orchard et al., 2015) and accumulate threshold-crossing activity, following the event-camera contrast mechanism (Lichtsteiner et al., 2008; Gallego et al., 2022). Activity is pooled onto an $H_R \times W_R$ spatial grid and accumulated into T_R ordered acquisition groups. Summing positive and negative activity gives

$$R_I^{\text{rot}}, R_E^{\text{rot}} \in \mathbb{R}_+^{T_R \times H_R \times W_R}. \tag{7}$$

The responses share a spatial grid and align acquisition groups by ordinal index. We compare their responses using whole-response ℓ_1 normalization and an elementwise square-root embedding:

$$S_{\text{resp}}^{\text{NCal}} = \cos\Big(\phi(R_I^{\xi}), \phi(R_E)\Big), \qquad \phi(R) = \sqrt{\frac{R}{\sum_u R_u + \varepsilon}}. \tag{8}$$

Cosine comparison of the flattened responses measures overlap between normalized activity distributions. Appendix B.3 gives the construction details.

mini N-ImageNet: spatial–polarity response. The acquisition protocol specifies planar translation (Kim et al., 2021). Per-recording phase is unspecified. We construct a response for each direction in $\mathcal{D} = \{+x, -x, +y, -y\}$ by accumulating consecutive log-intensity changes along its translation path. Positive and negative changes are pooled separately onto the spatial grid. Uniformly averaging the directional responses yields a phase-agnostic representation,

$$R_I^{\xi}, R_E \in \mathbb{R}_+^{|\mathcal{P}| \times H_R \times W_R}, \tag{9}$$

where $\mathcal{P}$ indexes negative and positive polarity channels. The observed events are accumulated over the recording into the same spatial–polarity layout. After whole-response ℓ_1 normalization, similarity is

$$\overline{R} = \frac{\operatorname{vec}(R)}{\|\operatorname{vec}(R)\|_1 + \varepsilon}, \qquad S_{\text{resp}}^{\text{miniNIN}} = \cos\Big(\overline{R}_I^{\xi}, \overline{R}_E\Big). \tag{10}$$

For controlled horizontal or vertical acquisition, the response averages the two opposite directions along the documented axis. All translational variants share the frozen display mapping, retained spatial region, pooling, and similarity rule. Appendix B.4 specifies these settings.

Static differential responses. The mini N-ImageNet test and controlled-motion comparisons use the same static differential response. Let I_{disp} be the grayscale image after the same training-calibrated display mapping used by the acquisition-informed response. We define

$$L_I = \log(\max(I_{\text{disp}}, \varepsilon_L)), \qquad G_I = \mathcal{P}_R([|\partial_x L_I|, |\partial_y L_I|]), \tag{11}$$

where ∂_x, ∂_y are Sobel derivatives and $\mathcal{P}_R$ sum-pools the shared retained region onto the $H_R \times W_R$ grid. For $E = \{(x_n, y_n, t_n, p_n)\}_{n=1}^{N_E}$, we accumulate all events into an unsigned count map and compute

$$H_E(x, y) = \sum_{n=1}^{N_E} \mathbf{1}[(x_n, y_n) = (x, y)], \qquad G_E = \mathcal{P}_R([|\partial_x H_E|, |\partial_y H_E|]). \tag{12}$$

The two channels retain horizontal and vertical derivative magnitudes. Both responses use whole-response ℓ_1 normalization and cosine comparison as in Eq. 10. The fixed-readout comparisons also use a simple spatial-activity score A, comparing pooled grayscale-image gradient magnitudes with pooled unsigned event counts. Appendix B.5 defines both static constructions.

Inference and reuse. For a fixed response specification, representations can be precomputed per observation; each shortlist then requires K similarities and an ordering rule.

3.3 A Structure-Aware Initial Ranker: PP-UOT

PP-UOT supplies structure-aware initial rankings for the controlled comparisons. Trained from scratch on image–event pairs, it represents each observation as sparse weighted local supports.

Image supports originate from an $H_S \times W_S$ spatial grid combining learned appearance features and deterministic structural cues. Event supports use the same grid and T_S ordered temporal bins; a leaky recurrent encoder processes local event statistics across time. Mass-guided selection retains K_I image supports and K_E event supports, forming the weighted measures

$$\mu_I = \sum_{i=1}^{K_I} \alpha_i \, \delta_{\zeta_i^I}, \qquad \mu_E = \sum_{j=1}^{K_E} \beta_j \, \delta_{\zeta_j^E}, \qquad \zeta = (x, c, f), \tag{13}$$

where x is spatial position, c a structural-mark distribution, f a learned feature, and δ_ζ a unit point mass at support ζ. The positive weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ each sum to one.

The pairwise cost matrix C combines discrepancies in position, structural marks, learned features, and support mass. Unbalanced optimal transport allows the matched mass to vary while penalizing deviations from the input support weights (Chizat et al., 2018). With entropic regularization (Cuturi, 2013), the ideal transport plan is

$$P^\star = \arg\min_{P \geq 0} \left[\langle P, C \rangle + \epsilon \mathcal{H}(P) + \rho D_{\mathrm{KL}}(P\mathbf{1} \| \boldsymbol{\alpha}) + \rho D_{\mathrm{KL}}(P^\top \mathbf{1} \| \boldsymbol{\beta}) \right], \tag{14}$$

where $\mathcal{H}$ is the negative-entropy regularizer and D_{KL} the generalized Kullback–Leibler divergence. The parameters ϵ and ρ weight regularization and marginal relaxation, respectively. A fixed 20-step generalized Sinkhorn solver yields the approximate plan $\widetilde{P} = P^{(20)}$, used in the correspondence score

$$S_{\mathrm{corr}}^{\text{PP-UOT}}(I, E) = - \left[\frac{\langle \widetilde{P}, C \rangle}{\|\widetilde{P}\|_1 + \varepsilon_0} + \lambda_{\mathrm{dev}} D_{\mathrm{marg}}(\widetilde{P}) \right], \tag{15}$$

where D_{marg} sums the ℓ_1 marginal deviations from $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. PP-UOT is trained using symmetric contrastive classification of paired observations in both retrieval directions. Appendix B.2 provides implementation and training details.

4 Experiments

4.1 Experimental Setup

We evaluate paired retrieval on N-Caltech101 (Orchard et al., 2015) and mini N-ImageNet (Kim et al., 2021). Avg R@1 averages image-to-event (I→E) and event-to-image (E→I) top-1 accuracy; accuracies are percentages and differences are percentage points (pp).

Primary, static-response, and fixed-fusion test results average three training runs. Comparisons hold each trained model and its top-five candidates fixed, preserving R@5 and R@10. Follow-up comparisons reuse previously evaluated test partitions with response constructions and score rules fixed using training and internal validation data.

On mini N-ImageNet, internal validation comes from the training pool; *Original Test* is the released mini original validation split. *Vertical* and *Horizontal* share its image identities with motion-specific event recordings.

Paired 95% confidence intervals (CIs) use $10{,}000$ bootstrap resamples of paired identities, conditional on the evaluated checkpoints (Appendix C).

4.2 Improving Initial Rankings with Response Verification

We first test whether acquisition-informed response-only reranking (Sec. 3.2) improves the initial test-set rankings (Table 1).

All six mean changes are positive, ranging from $+0.643$ to $+18.243$ pp; five paired confidence intervals exclude zero. With models and candidates fixed, response comparison can resolve pairing ambiguities left by different learned correspondence models.

CEIA on mini N-ImageNet improves E→I by 4.653 pp and reduces I→E by 3.367 pp; its bidirectional interval includes zero. This directional difference motivates examining the response representation in Sec. 4.3 and its combination with correspondence in Sec. 4.4. Appendix D reports complete directional, seed-level, and cross-split results.

4.3 Response Representations and Candidate Discrimination

Spatial organization. Summing responses over spatial positions reduces three-run PP-UOT validation Avg R@1 from 87.551% to 31.333% on N-Caltech101 and from 84.283% to 18.077% on mini N-ImageNet. These reductions show the value of retaining spatial activity locations. Appendix G reports full controls and sensitivity analyses.

Table 1: **Acquisition-informed verification of initial test-set rankings.** RV denotes acquisition-informed response-only verification of the same top-five candidates. Recall values are mean ± population standard deviation across three runs (%). Δ gives the Avg R@1 change (pp), with a paired 95% CI. Pretraining is inherited by RV.

Model	Pretraining	E→I R@1	I→E R@1	Avg R@1	Δ	95% CI
N-Caltech101						
EventBind	CLIP	67.226 ± 0.689	69.997 ± 1.223	68.612 ± 0.732	–	–
+ RV		84.821 ± 0.554	88.889 ± 0.451	**86.855 ± 0.394**	**+18.243**	[16.527, 19.959]
CEIA	CLIP	77.778 ± 0.908	78.871 ± 0.755	78.324 ± 0.804	–	–
+ RV		87.719 ± 0.347	92.220 ± 0.408	**89.969 ± 0.325**	**+11.645**	[10.005, 13.285]
PP-UOT	None	82.787 ± 2.136	83.346 ± 2.463	83.066 ± 2.264	–	–
+ RV		85.177 ± 0.530	84.643 ± 0.743	**84.910 ± 0.629**	**+1.843**	[0.242, 3.445]
mini N-ImageNet						
EventBind	CLIP	66.240 ± 0.564	68.087 ± 0.423	67.163 ± 0.371	–	–
+ RV		79.787 ± 0.124	73.973 ± 0.327	**76.880 ± 0.199**	**+9.717**	[8.783, 10.643]
CEIA	CLIP	82.507 ± 0.131	83.413 ± 0.398	82.960 ± 0.193	–	–
+ RV		**87.160 ± 0.257**	80.047 ± 0.290	83.603 ± 0.034	+0.643	[-0.240, 1.520]
PP-UOT	None	40.687 ± 1.293	41.320 ± 0.653	41.003 ± 0.917	–	–
+ RV		54.553 ± 1.566	52.353 ± 0.160	**53.453 ± 0.823**	**+12.450**	[11.723, 13.173]

Acquisition-group correspondence on N-Caltech101. Keeping the eight acquisition maps separate yields 87.551% three-run mean validation accuracy, compared with 86.149% after combining them at the same spatial locations. The 1.402 pp difference measures the added value of retaining acquisition groups once spatial layout is preserved (Appendix G.1).

A complementary fixed-model intervention changes only the correspondence between predicted and observed acquisition groups. The documented motion-based order exceeds all $1{,}000$ sampled random reorderings, while a tested adjacent swap remains competitive (Fig. 2). Together, these controls support useful coarse acquisition-group organization with tolerance to a local change in order.

Static alternatives on mini N-ImageNet test data. Table 2 compares acquisition-informed similarity S with static differential similarity G (Sec. 3.2) on Original Test, using the same candidates and calibrated spatial preprocessing for each EventBind and CEIA checkpoint.

Table 2: **Static versus acquisition-informed responses on mini N-ImageNet Original Test.** S-only and G-only are three-run Avg R@1 means (%). $G-S$ is the paired three-run mean ± population standard deviation (pp); directional columns give the corresponding three-run mean differences. The 95% CI is for bidirectional $G - S$.

Model	S-only	G-only	Avg $G - S$	I→E	E→I	95% CI
EventBind	76.880	**80.453**	$+3.573 \pm 0.029$	+7.047	+0.100	[3.137, 4.017]
CEIA	83.603	**88.210**	$+4.607 \pm 0.077$	+8.520	+0.693	[4.130, 5.080]

G exceeds S in all six model–seed bidirectional comparisons, and both three-run mean differences have positive paired intervals. The directional means show that most of this advantage occurs in I→E. For CEIA in this direction, changing S to G raises accuracy from 80.047% to 88.567%, above the initial 83.413%. Thus, the response representation itself changes verification's effect on the same candidates. Appendix E gives the complete per-seed and directional results.

Added and reused response scores. On Vertical and Horizontal, S uses the documented motion axis. We first add S or G to the same background combining correspondence and simple spatial activity; after including G, we compare adding S with reusing G under the same fixed rule. Reuse controls for gains from another use of existing static evidence (Appendix C.5).

Choosing G gives higher accuracy in all four comparisons, with positive paired intervals (Fig. 3). After G, adding S improves PP-UOT's horizontal accuracy by 0.110 pp (95% CI $[0.013, 0.210]$).

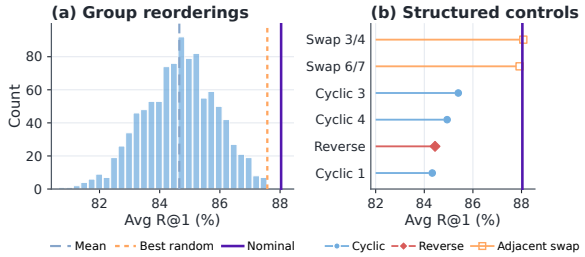


Figure 2: **Acquisition-group correspondence on N-Caltech101.** PP-UOT seed-42 validation results. The nominal order exceeds all 1,000 sampled random reorderings, while selected local reorderings remain competitive. Observed responses and candidate sets are fixed.

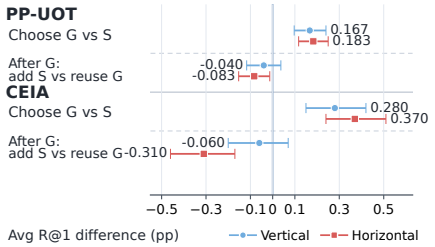


Figure 3: **Controlled-motion response comparisons.** Avg R@1 differences (pp) with paired 95% CIs. Positive "Choose G vs. S" values favor G; negative "After G" values favor reusing G. PP-UOT averages three runs; CEIA uses seed 42.

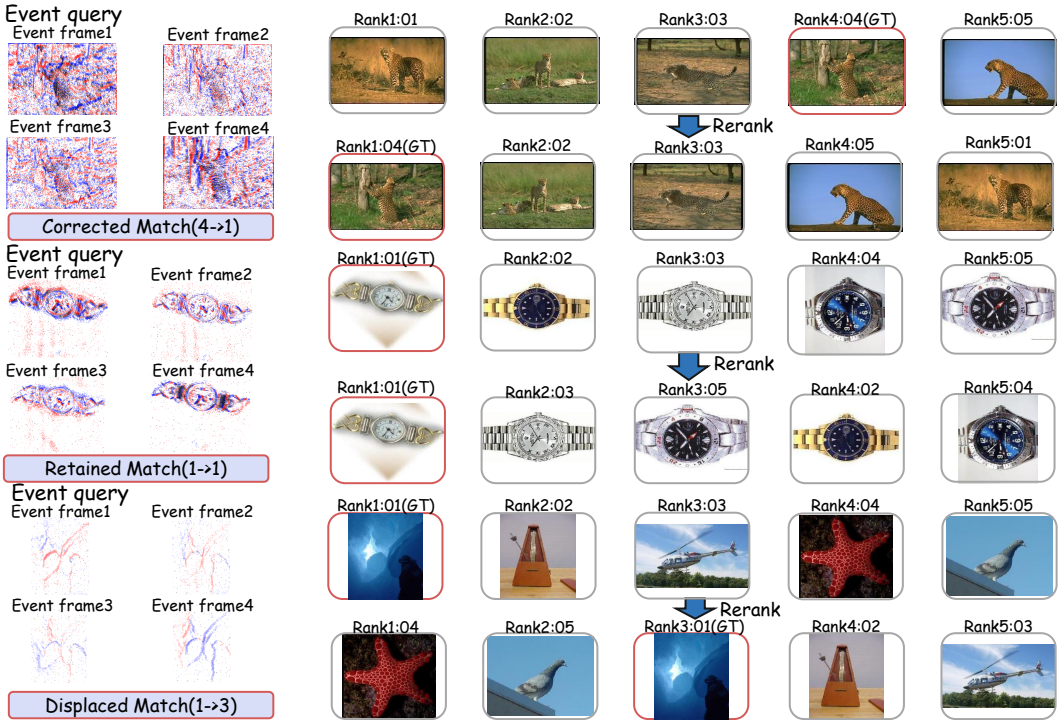


Figure 4: **Qualitative event-to-image reranking examples.** PP-UOT examples show corrected ($4 \to 1$), retained ($1 \to 1$), and displaced ($1 \to 3$) paired matches, from top to bottom. Each event query is displayed in four temporal frames, ordered left to right and then top to bottom. Upper and lower image rows show initial and reranked candidates. Red borders and GT identify the paired image; candidate IDs remain fixed within each example. All five candidate identities are preserved.

Reusing G yields the higher mean in all four settings; intervals favor reuse on Horizontal and include zero on Vertical. Appendix E gives endpoints and additional controls.

4.4 Combining Correspondence and Response Scores

We test whether retaining the correspondence score improves retrieval over response-only reranking.

From validation to fixed test fusion. The correspondence weight $\lambda = 0.25$ gives the highest tested mean accuracy on both PP-UOT validation splits (Appendix F.2). We hold Eq. 5 at this weight across the eight test settings in Table 3.

Table 3: **Fixed-weight test fusion and a representative directional analysis.** C: correspondence; X: acquisition-informed S or static G; $C + X$: fusion at $\lambda = 0.25$. **(a)** Three-run Avg R@1 (%) and gain over X-only (pp). **(b)** CEIA on mini N-ImageNet Original Test, I→E: three-run R@1 and fusion-minus-response gain decomposition. Decomposition terms use all queries as the denominator.

(a) Fixed-weight fusion across settings

Model	X	C	X-only	$C + X$	Δ
N-Caltech101					
EventBind	S	68.612	86.855	**89.423**	+2.568
CEIA	S	78.324	89.969	**92.919**	+2.949
PP-UOT	S	83.066	84.910	**90.936**	+6.026
mini N-ImageNet Original Test					
EventBind	S	67.163	76.880	**80.463**	+3.583
EventBind	G	67.163	80.453	**82.773**	+2.320
CEIA	S	82.960	83.603	**89.483**	+5.880
CEIA	G	82.960	88.210	**92.010**	+3.800
PP-UOT	S	41.003	53.453	**55.050**	+1.597

(b) CEIA: mini I→E

Rule	S	G
Initial C	83.413	83.413
X-only	80.047	88.567
$C + X$	**87.767**	**92.073**
Contribution	S	G
Displacement reduction	+8.107	+4.027
Correction change	−0.387	−0.520
Net gain	**+7.720**	**+3.507**

Across all eight settings, including the strong static responses, fusion exceeds both endpoint rankings and improves Avg R@1 over response-only ranking by 1.597–6.026 pp. Both gains hold in every evaluated training run and direction; all 16 paired 95% intervals for three-run bidirectional gains lie above zero. Appendix F.3 reports the complete intervals and across-seed variation.

Understanding the CEIA directional loss. A *rescue* moves the paired match from ranks $2, \ldots, K$ to rank 1; a *damage* displaces an initially correct first-ranked match. The selected PP-UOT examples in Fig. 4 illustrate correction, retention, and displacement within the same candidate set. Correction and displacement counts determine the net accuracy change:

$$\Delta \mathrm{R@1}_{\mathrm{pp}} = 100 \frac{N_{\mathrm{rescued}} - N_{\mathrm{damaged}}}{N_{\mathrm{queries}}}. \tag{16}$$

For CEIA mini I→E, fusion with the same S raises accuracy from 80.047% to 87.767%, above the initial 83.413%. In Table 3(b), fewer displaced correct matches contribute +8.107 pp, while the change in successful corrections contributes −0.387 pp, producing a net +7.720 pp over S-only. The G branch shows the same pattern, with a net +3.507 pp gain. Appendix F.4 gives both directions and the complete correction counts. Response choice affects candidate discrimination in Sec. 4.3; retaining correspondence affects how the selected response changes the initial ranking.

5 Conclusion

We studied response verification for paired image–event retrieval across three models and two controlled benchmarks. Response comparisons show useful coarse acquisition organization on N-Caltech101 and stronger static verification for EventBind and CEIA on mini N-ImageNet. A validation-selected fusion rule improves test accuracy over the initial and corresponding response-only rankings across the evaluated settings. The CEIA analysis links this improvement to fewer displaced correct matches. These findings motivate evaluating verification representations together with score integration, and extending the study to natural recordings with free camera motion, dynamic scenes, and varying illumination.

AI USE STATEMENT

We used generative AI tools, including ChatGPT, to assist with implementation and analysis code, manuscript and LaTeX editing, literature review, mathematical exposition, experimental interpretation, and feedback on research framing and experimental design. AI-assisted drafts also informed the preliminary figure layout; the final method diagram was manually redrawn in PowerPoint and exported as PDF. The authors reviewed and revised AI-assisted material and checked reported values against experimental records. All quantitative results were obtained from the training and evaluation pipelines described in this paper. The authors take responsibility for the complete submission.

ETHICS STATEMENT

Our experiments use existing N-Caltech101 and mini N-ImageNet benchmark pairs acquired from displayed source images under controlled camera motion (Orchard et al., 2015; Kim et al., 2021). The evaluation focuses on paired-instance retrieval. Applications to natural recordings require attention to privacy, consent, and performance across acquisition conditions. Reuse and redistribution of benchmark data and pretrained models should follow the conditions specified by their providers.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B specify the retrieval interface, response constructions, and initial ranker. Appendix C documents data partitions, calibration, training settings, fixed candidate sets, score readouts, and paired bootstrap inference. Appendices D–G provide seed-level results, controlled comparisons, the fusion sweep, fixed-weight test fusion, and sensitivity analyses. Seed-wise statistics describe variation across evaluated checkpoints, and paired bootstrap intervals summarize query-level uncertainty conditional on those checkpoints.

REFERENCES

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *Computer Vision – ECCV 2020*, pp. 136–152, 2020.

Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Computer Vision – ECCV 2020*, pp. 726–743, 2020.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal image-TExt representation learning. In *Computer Vision – ECCV 2020*, pp. 104–120, 2020.

Zhiwen Chen, Junhui Hou, Zhiyu Zhu, Jinjian Wu, and Guangming Shi. Scaling dense event-stream pretraining from visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8011–8022, 2026.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the IEEE 11th International Conference on Computer Vision*, 2007.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *International Conference on Learning Representations*, 2025.

Guillermo Gallego, Jon E. A. Lund, Elias Mueggler, Henri Rebecq, Tobi Delbruck, and Davide Scaramuzza. Event-based, 6-DOF camera tracking from photometric depth maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2402–2412, 2018a.

Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3867–3876, 2018b.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5633–5643, 2019.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Cheng Gu, Erik Learned-Miller, Daniel Sheldon, Guillermo Gallego, and Pia Bideau. The spatio-temporal poisson point process: A simple model for the alignment of event camera data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13495–13504, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Sungheon Jeong, Hanning Chen, Sanggeon Yun, Suhyeon Cho, Wenjun Huang, Xiangjian Liu, and Mohsen Imani. Cross-modal event encoder: Bridging image-text knowledge to event streams. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3213–3222, 2026.

Xiang Ji, Jiaxin Wei, Yifu Wang, Huiliang Shang, and Laurent Kneip. Cross-modal place recognition in image databases using event-based sensors. *arXiv preprint arXiv:2307.01047*, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025.

Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2146–2156, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Computer Vision – ECCV 2018*, pp. 212–228, 2018.

Quanmin Liang, Qiang Li, Shuai Liu, Xinzi Cao, Jinyi Lu, Feidiao Yang, Wei Zhang, Kai Huang, and Yonghong Tian. Efficient event camera data pretraining with adaptive prompt fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8656–8667, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022.

Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Tien Manh Luong, Khai Nguyen, Nhat Ho, Reza Haffari, Dinh Phung, and Lizhen Qu. Revisiting deep audio-text retrieval through the lens of transportation. In *International Conference on Learning Representations*, 2024.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3476–3485, 2017.

Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015.

Federico Paredes-Vallés and Guido C. H. E. de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3446–3455, 2021.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. In *Proceedings of the 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 969–982. PMLR, 2018.

Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3857–3866, 2019.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, 2025.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1731–1740, 2018.

Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pp. 1470–1477, 2003.

Timo Stoffregen and Lindsay Kleeman. Event cameras, contrast maximization and reward functions: An analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, pp. 776–794, 2020.

Andreas Veit, Serge J. Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1781–1789, 2017.

Ziyi Wu, Xudong Liu, and Igor Gilitschenski. EventCLIP: Adapting CLIP for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023.

Wenhao Xu, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. CEIA: CLIP-based event-image alignment for open-world event-based understanding. In *Computer Vision – ECCV 2024 Workshops*, volume 15646 of *Lecture Notes in Computer Science*, pp. 1–18. Springer, 2025.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10699–10709, 2023.

Yan Yang, Liyuan Pan, Dongxu Li, and Liu Liu. EZSR: Event-based zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4628–4638, 2025.

Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. EventBind: Learning a unified representation to bind them all for event-based open-world understanding. In *Computer Vision – ECCV 2024*, pp. 477–494, 2024.

Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5819–5828, 2024.

A Extended Related Work

We expand Section 2 through four directions that motivate the correspondence and response comparisons in this study.

A.1 Multimodal Alignment, Geometry, and Multiple Similarity Criteria

Multimodal learning constructs comparable representations for cross-modal matching. Contrastive Multiview Coding learns across views (Tian et al., 2020); Contrastive Language–Image Pretraining (CLIP) and ALIGN scale image–text alignment (Radford et al., 2021; Jia et al., 2021). FLAVA combines language and vision learning, and ImageBind extends alignment to further modalities (Singh et al., 2022; Girdhar et al., 2023). These representations support similarity-based transfer and retrieval.

Studies of embedding geometry characterize modality gaps, modality-dependent organization, object bias, and information imbalance (Liang et al., 2022; Schrodi et al., 2025; Eslami & de Melo, 2025). Dufumier et al. (2025) examine the multimodal interactions captured by contrastive learning. VLM2Vec uses task instructions for retrieval embeddings (Jiang et al., 2025), and Conditional Similarity Networks learn subspaces for different visual criteria (Veit et al., 2017). Together, these studies motivate comparing observations through multiple representations. We evaluate a fixed response space after learned image–event correspondence, measuring both its effect on candidate order and the benefit of combining the two scores. Both comparisons can exploit shared scene structure.

A.2 Event Representations and Image–Event Knowledge Transfer

Event representations organize asynchronous brightness changes for learning and matching (Gallego et al., 2022). Histograms of Averaged Time Surfaces (HATS) summarize local temporal context (Sironi et al., 2018); differentiable event grids and recurrent surfaces learn structured features (Gehrig et al., 2019; Cannici et al., 2020). N-ImageNet provides a large-scale recognition benchmark with controlled acquisition variants (Kim et al., 2021).

Paired observations enable image–event knowledge transfer through event pretraining (Yang et al., 2023), alongside sequence modeling with state-space models (Zubic et al., 2024). EventCLIP converts event counts into frames for CLIP-based recognition and learns lightweight feature adapters in the few-shot setting (Wu et al., 2023). EventBind learns a unified image–text–event representation (Zhou et al., 2024), and CLIP-based Event–Image Alignment (CEIA) connects event embeddings to CLIP image features (Xu et al., 2025). Further work studies zero-shot recognition, adaptive prompt fusion, and dense pretraining from visual foundation models (Yang et al., 2025; Liang et al., 2025; Chen et al., 2026). Cross-Modal Event Encoder transfers image–text knowledge through event–image alignment and evaluates retrieval (Jeong et al., 2026). These approaches supply the representation-learning context for our initial rankers. We use fixed EventBind and CEIA candidates to compare acquisition-informed and static responses, making the spatial and temporal structure used in the additional verification step explicit.

A.3 Structured Cross-Modal Matching, Reranking, and Verification

Structured matching connects local parts across modalities: Stacked Cross Attention Networks (SCAN) model region–word correspondence (Lee et al., 2018), UNITER uses word–region optimal transport (Chen et al., 2020), and partial transport supports audio–text retrieval with misaligned training pairs (Luong et al., 2024). Our Point-Pattern Unbalanced Optimal Transport (PP-UOT) ranker uses weighted local supports to provide controlled correspondence rankings.

Image retrieval combines initial ranking with specialized candidate comparison. Video Google uses visual words and local spatial matching (Sivic & Zisserman, 2003); spatial verification reranks large-vocabulary results (Philbin et al., 2007), and verified matches support query expansion (Chum et al., 2007). DELF learns attentive local descriptors (Noh et al., 2017), while DELG combines local and global retrieval features (Cao et al., 2020).

For event-to-image place recognition, Ji et al. (2023) retrieve image candidates and rerank them with a learned similarity classifier. The classifier combines image and event features through compact

bilinear pooling and a multilayer perceptron; the retrieval and classification branches are trained jointly. Their evaluation on Brisbane-Event-VPR targets geographic place matches.

We study one-to-one image–event instance retrieval using deterministic response mappings. With trained rankers and their candidate sets held fixed, we compare acquisition-informed and static responses and test whether combining each response score with learned correspondence improves on response-only ranking. This design lets us assess response construction and score integration separately within the same retrieval setting.

A.4 Event Formation, Motion Compensation, and Physical Consistency

Event-camera pixels respond to log-intensity changes (Lichtsteiner et al., 2008; Gallego et al., 2022), supporting explicit rendering and measurement models. ESIM synthesizes events through scene rendering (Rebecq et al., 2018). Contrast maximization warps events under candidate motion models for motion, depth, and optical-flow estimation (Gallego et al., 2018b); related analysis studies its reward functions (Stoffregen & Kleeman, 2019). Direct event-based tracking compares photometric scene information with camera motion (Gallego et al., 2018a).

Spatial and photometric structure also supports video reconstruction (Rebecq et al., 2019) and Poisson-process event alignment (Gu et al., 2021). Photometric constancy supports self-supervised image reconstruction from events (Paredes-Vallés & de Croon, 2021). We apply the appearance–motion relationship to compact retrieval responses: images are rendered along a fixed motion specification and brightness-change activity is pooled for candidate comparison. Acquisition-group controls on N-Caltech101 and static/repeated-score comparisons on mini N-ImageNet measure which retained structures contribute to retrieval under the evaluated acquisition conditions.

B Detailed Method and Implementation

We specify the learned initial ranker (Stage I), fixed response mappings, and within-shortlist verification scores (Stage II).

B.1 Candidate-Constrained Response Verification

The initial ranking and final within-shortlist permutation follow Eqs. 1 and 3. For an event query, candidate r is scored as $S_{\mathrm{resp}}(I_r, E_q)$; for an image query, it is scored as $S_{\mathrm{resp}}(I_q, E_r)$. The primary protocol uses $K = 5$ as a common candidate budget across models and datasets. Appendix G.5 evaluates $K \in \{1, 3, 5, 10\}$ on internal validation. Reranking preserves shortlist membership and all lower-ranked positions. Consequently, writing recall as a fraction,

$$\mathrm{R@}L_{\mathrm{final}} = \mathrm{R@}L_{\mathrm{corr}} \quad (L \geq K), \tag{17}$$

$$\mathrm{R@}1_{\mathrm{final}} \leq \mathrm{R@}K_{\mathrm{corr}}. \tag{18}$$

Appendix F.1 relates top-1 changes to corrections and displaced correct matches.

Responses of dimension D_R can be precomputed per observation, giving $\mathcal{O}(KD_R)$ comparison cost after lookup. In the follow-up tests, exact response-only and fused-score ties retain initial order. Fusion uses Eq. 5 to define the same candidate-constrained permutation and inherits the bounds above.

B.2 Full PP-UOT Formulation

PP-UOT learns correspondence between sparse weighted local supports using unbalanced optimal transport and the loss in Eq. 50. The trained ranker supplies fixed candidates for response verification.

Image support construction. Given an RGB image I, we resize it to the common input resolution and form local candidates on an $H_S \times W_S$ spatial lattice. Let I_{gray} denote grayscale intensity and

let g_x, g_y be 3×3 Sobel responses. The gradient quantities and deterministic five-channel prior are

$$g = \sqrt{g_x^2 + g_y^2 + \varepsilon_g}, \qquad \theta = \operatorname{atan2}(g_y, g_x), \qquad \varepsilon_g = 10^{-6}, \tag{19}$$

$$r_I = [g,\, \bar{g},\, q,\, \sin\theta,\, \cos\theta]\,, \tag{20}$$

$$\bar{g} = \mathrm{AvgPool}_{3\times3}(g), \qquad q = \left|I_{\mathrm{gray}} - \mathrm{AvgPool}_{5\times5}(I_{\mathrm{gray}})\right|. \tag{21}$$

A shallow convolutional encoder produces learned appearance features, which are pooled to the same lattice and combined with these deterministic cues. A shared local projection produces h_i^I. Prediction heads output structural-mark logits ℓ_i^I, local matching feature f_i^I, and positive transport mass m_i^I:

$$c_i^I = \mathrm{softmax}(\ell_i^I), \qquad \zeta_i^I = (x_i^I, c_i^I, f_i^I), \qquad m_i^I > 0. \tag{22}$$

The image lattice fixes x_i^I.

Event support construction and temporal encoding. PP-UOT consumes every event in the complete recording and preserves stable timestamp order throughout the temporal-bin construction. Within each recording, timestamps and temporal-bin indices are

$$u_n = \frac{t_n - t_{\min}}{\max(t_{\max} - t_{\min}, \varepsilon_t)}, \qquad b_n = \min\{T_S - 1, \lfloor T_S u_n \rfloor\}, \tag{23}$$

where $\varepsilon_t = 1$ in the stored timestamp unit. We use $T_S = 8$ ordered temporal bins over an $H_S \times W_S = 8 \times 8$ spatial lattice, giving 512 time–space candidate cells before sparsification. For each local spatiotemporal cell, let N, N_+, N_- be total and polarity-specific event counts. The count normalization and local event-statistics vector are

$$M_N = \max_{t,s} N_{t,s}, \qquad \widehat{N} = \frac{N}{M_N}, \qquad \widehat{N}_\pm = \frac{N_\pm}{M_N}, \tag{24}$$

$$v = \left[\widehat{N},\, \widehat{N}_+,\, \widehat{N}_-,\, b_{\mathrm{pol}},\, \Delta t,\, \bar{x},\, \bar{y}\right], \tag{25}$$

$$b_{\mathrm{pol}} = \begin{cases} \dfrac{N_+ - N_-}{N}, & N > 0, \\ 0, & N = 0, \end{cases} \qquad \Delta t = \begin{cases} \dfrac{u_{\max} - u_{\min}}{N - 1}, & N > 1, \\ 0, & N \le 1. \end{cases} \tag{26}$$

Here $(\bar{x}, \bar{y})$ is the mean normalized event position inside the cell. Only cells containing observed event activity are eligible to become transport supports.

For each spatial location s, the ordered temporal sequence is processed by a leaky recurrent encoder with a differentiable spike-like response (Neftci et al., 2019). Its activity gate, membrane update, spike-like response, and temporal representation are

$$a_{t,s} = \mathbb{I}[N_{t,s} > 0], \qquad d_{t,s} = W_{\mathrm{in}} v_{t,s} + b_{\mathrm{in}}, \qquad \ell = \sigma(\ell_{\mathrm{raw}}), \tag{27}$$

$$m_{t,s} = \ell \odot m_{t-1,s} + a_{t,s} d_{t,s}. \tag{28}$$

$$s_{t,s} = a_{t,s}\sigma\left(\frac{m_{t,s} - \vartheta}{\tau_{\mathrm{spk}}}\right), \qquad o_{t,s} = W_o[m_{t,s}; s_{t,s}] + b_o. \tag{29}$$

The membrane state is initialized at the start of each recording and carried forward across all temporal bins. For each active cell, we form

$$z_{t,s} = [v_{t,s}, o_{t,s}, \tau_t, \chi_s], \tag{30}$$

where τ_t is normalized temporal-bin position and χ_s the coarse spatial coordinate. A shared projection maps $z_{t,s}$ to latent representation h_j^E.

Prediction heads produce

$$c_j^E = \mathrm{softmax}(\ell_j^E), \qquad q_j = \mathrm{softplus}(W_q h_j^E + b_q), \tag{31}$$

$$x_j^E = \mathrm{clip}_{[0,1]}\left(x_j^{\mathrm{grid}} + \gamma_x \tanh(W_x h_j^E + b_x)\right). \tag{32}$$

The resulting support descriptor is $\zeta_j^E = (x_j^E, c_j^E, f_j^E)$.

Mass-guided sparse support selection. Image transport importance is the predicted positive mass m_i^I. Given event strengths q_j and support budgets K_I, K_E, event masses, hard support sets, and normalized retained masses are

$$m_j^E = q_j \left(1 + \frac{N_j}{M_N}\right). \tag{33}$$

$$\mathcal{S}_I = \mathrm{TopK}_{K_I}(\{m_i^I\}), \qquad \mathcal{S}_E = \mathrm{TopK}_{K_E}(\{m_j^E : N_j > 0\}), \tag{34}$$

$$\alpha_i = \frac{m_i^I}{\sum_{r\in\mathcal{S}_I} m_r^I}, \qquad \beta_j = \frac{m_j^E}{\sum_{r\in\mathcal{S}_E} m_r^E}. \tag{35}$$

During training, the forward pass always uses hard Top-K_{sel} IDs obtained from the raw positive masses, where K_{sel} is K_I or K_E for the corresponding modality. A successive-softmax straight-through surrogate propagates gradients through the mass-prediction pathway. For valid candidate set $\mathcal{V}$,

$$\widetilde{\ell}_r = \frac{\log\max(m_r, 10^{-8}) - \mu_{\log m}}{\sigma_{\log m} + 10^{-8}}, \qquad r \in \mathcal{V}. \tag{36}$$

Initialize $q_r^{(0)} = 0$ and, for selection step $s = 1, \ldots, K_{\mathrm{sel}}$, compute

$$u_r^{(s)} = \max(1 - q_r^{(s-1)}, 10^{-6}), \tag{37}$$

followed by

$$p_r^{(s)} = \mathrm{Softmax}_{r\in\mathcal{V}}\left(\frac{\widetilde{\ell}_r}{\tau_{\mathrm{sel}}} + \log u_r^{(s)}\right), \qquad q_r^{(s)} = q_r^{(s-1)} + p_r^{(s)}. \tag{38}$$

The surrogate satisfies $\sum_{r\in\mathcal{V}} q_r^{(K_{\mathrm{sel}})} = K_{\mathrm{sel}}$. If

$$h_r = \mathbb{I}[r \in \mathrm{TopK}_{K_{\mathrm{sel}}}(m)] \tag{39}$$

is the hard forward indicator, the straight-through gate is

$$g_r^{\mathrm{ST}} = h_r - \mathrm{stopgrad}(q_r^{(K_{\mathrm{sel}})}) + q_r^{(K_{\mathrm{sel}})} \tag{40}$$

following the straight-through principle (Bengio et al., 2013). We use $\tau_{\mathrm{sel}} = 1.0$ for both datasets. The straight-through gate acts only on selected masses; coordinates, marks, features, and event-side temporal representations are gathered using the hard support IDs. Inference therefore always uses the deterministic hard support set.

Structural cost and unbalanced transport. The selected supports define

$$\mu_I = \sum_{i\in\mathcal{S}_I} \alpha_i \delta_{\zeta_i^I}, \qquad \mu_E = \sum_{j\in\mathcal{S}_E} \beta_j \delta_{\zeta_j^E}. \tag{41}$$

For retained image support i and event support j, the pairwise cost is

$$C_{ij} = \lambda_x d_x(i,j) + \lambda_c d_c(i,j) + \lambda_f d_f(i,j) + \lambda_m d_m(i,j), \tag{42}$$

where

$$\begin{aligned} d_x(i,j) &= \|x_i^I - x_j^E\|_2, \quad & d_c(i,j) &= \mathrm{JSD}(c_i^I, c_j^E), \\ d_f(i,j) &= \|\widehat{f}_i^I - \widehat{f}_j^E\|_2^2, \quad & d_m(i,j) &= |\alpha_i - \beta_j|, \end{aligned} \tag{43}$$

with $\widehat{f} = f/\max(\|f\|_2, \varepsilon_f)$ and JSD denoting Jensen–Shannon divergence. The frozen weights are

$$(\lambda_x, \lambda_c, \lambda_f, \lambda_m) = (0.5, 1, 0.5, 0.1). \tag{44}$$

Event timing influences the support attributes through the temporal encoder.

We use the negative-entropy regularizer and generalized KL divergence

$$\mathcal{H}(P) = \sum_{i,j} P_{ij}(\log P_{ij} - 1), \tag{45}$$

$$D_{\mathrm{KL}}(r\|a) = \sum_i \left[r_i \log \frac{r_i}{a_i} - r_i + a_i \right], \tag{46}$$

with $0 \log 0 = 0$ and the continuous extension at $r_i = 0$. The reference masses a_i are positive. The linear mass terms are retained because the total transported mass varies during optimization. The transport objective combines entropic regularization (Cuturi, 2013) with relaxed marginal penalties (Chizat et al., 2018):

$$P^\star = \arg\min_{P \geq 0} \left[\langle P, C \rangle + \epsilon \mathcal{H}(P) + \rho D_{\mathrm{KL}}(P\mathbf{1}\|\boldsymbol{\alpha}) + \rho D_{\mathrm{KL}}(P^\top \mathbf{1}\|\boldsymbol{\beta}) \right]. \tag{47}$$

In practice, a fixed 20-step generalized Sinkhorn solver (Chizat et al., 2018; Peyré & Cuturi, 2019) produces the approximate plan $\widetilde{P} = P^{(20)}$, with $\epsilon = 0.2$ and $\rho = 1$. All training and retrieval scores are evaluated with this fixed-iteration plan:

$$S_{\mathrm{corr}}^{\text{PP-UOT}}(I, E) = -\left[\frac{\langle \widetilde{P}, C \rangle}{\|\widetilde{P}\|_1 + \varepsilon_0} + \lambda_{\mathrm{dev}} D_{\mathrm{marg}}(\widetilde{P}) \right], \tag{48}$$

where

$$D_{\mathrm{marg}}(P) = \|P\mathbf{1} - \boldsymbol{\alpha}\|_1 + \|P^\top \mathbf{1} - \boldsymbol{\beta}\|_1. \tag{49}$$

Bidirectional retrieval training. For a minibatch of paired observations, the score matrix has entries $S_{nm} = S_{\mathrm{corr}}^{\text{PP-UOT}}(I_n, E_m)$. PP-UOT uses symmetric bidirectional contrastive classification,

$$\mathcal{L}_{\text{PP-UOT}} = \frac{1}{2} \left[\mathrm{CE}\left(\frac{S}{\tau_{\mathrm{con}}}, \mathbf{y} \right) + \mathrm{CE}\left(\frac{S^\top}{\tau_{\mathrm{con}}}, \mathbf{y} \right) \right], \tag{50}$$

where CE denotes cross-entropy and $\mathbf{y}$ contains the diagonal correspondence labels. The loss updates the initial ranker through its paired correspondence scores; the resulting checkpoint is frozen before response verification.

Table 4: **PP-UOT implementation configuration.** Settings are shared across datasets unless explicitly separated.

Setting	Symbol	Value
Input resolution	$H_0 \times W_0$	64×64
Spatial support grid	$H_S \times W_S$	8×8
Temporal bins	T_S	8
Smoothed-gradient pooling	k_g	3
Local-contrast pooling	k_q	5
Structural-mark dimension	d_c	12
Local feature dimension	d_f	32
Temporal hidden dimension	d_h	64
Soft-spike temperature	τ_{spk}	0.35
Coordinate correction bound	γ_x	0.1
Image/event supports, N-Cal	(K_I, K_E)	$(32, 40)$
Image/event supports, mini	(K_I, K_E)	$(48, 60)$
Selection temperature	τ_{sel}	1.0
Structural cost weights	$(\lambda_x, \lambda_c, \lambda_f, \lambda_m)$	$(0.5, 1, 0.5, 0.1)$
Entropic regularization	ϵ	0.2
Marginal relaxation	ρ	1
Transport iterations	N_{OT}	20
Marginal-deviation weight	λ_{dev}	0.2
Contrastive temperature	τ_{con}	0.035

B.3 N-CALTECH101 ACQUISITION-INFORMED RESPONSE

N-Caltech101 pairs displayed source images with event recordings captured under controlled saccadic sensor motion (Orchard et al., 2015). The frozen verifier uses a retrieval-oriented rotational response that preserves spatial organization and coarse acquisition progression.

Geometry and shared acquisition progress. The documented motor path (Orchard et al., 2015) is represented by

$$(-0.5^\circ, +0.5^\circ) \to (0^\circ, -0.5^\circ) \to (+0.5^\circ, +0.5^\circ) \to (-0.5^\circ, +0.5^\circ). \tag{51}$$

Motor x and y map to camera yaw and pitch, respectively. For motor state $p = (p_x, p_y)$,

$$C(p) = R_y(p_x)R_x(p_y), \qquad R_{w\to c}(p) = C(p)^\top C(p_0), \tag{52}$$

with frozen rasterization matrix

$$K = \begin{bmatrix} 160 & 0 & 151.5 \\ 0 & 160 & 119.5 \\ 0 & 0 & 1 \end{bmatrix}. \tag{53}$$

The value 160 px is the effective angular-to-raster scale calibrated on training data for this rendering model.

Progress along each trajectory segment follows a shared event-rate cumulative distribution function (CDF) estimated from training recordings. Training event timestamps are normalized to $[0, 1]$, accumulated into 96 temporal microbins, divided into three consecutive 32-bin segments, and summed over the training recordings. Within segment s, cumulative normalized counts define a shared 33-point progress table F_s. For segment endpoints p_s, p_{s+1},

$$p_{s,k} = p_s + F_s(k)(p_{s+1} - p_s). \tag{54}$$

The same progress tables control interpolation speed along the documented path for every candidate image. The resulting response uses 97 rendered states and 96 state transitions.

Persistent-reference threshold response. For state k,

$$H_k = KR_{w\to c}(p_k)K^{-1}, \qquad I_k = \mathrm{Warp}(I, H_k), \qquad L_k = \log(\max(I_k, \varepsilon_L)), \tag{55}$$

with $\varepsilon_L = 1/255$. Let M_k be the warped stimulus-validity mask and

$$\nu_k = \mathbb{I}[M_k > \tau_M], \qquad V_k = \nu_{k-1}\nu_k, \qquad \tau_M = 0.999. \tag{56}$$

A persistent log-intensity reference is initialized as $\widehat{L}_0 = L_0$. For state k,

$$D_k = L_k - \widehat{L}_{k-1}. \tag{57}$$

The positive and negative threshold-crossing counts are

$$A_k^+ = V_k \left\lfloor \frac{[D_k]_+}{\kappa} \right\rfloor, \qquad A_k^- = V_k \left\lfloor \frac{[-D_k]_+}{\kappa} \right\rfloor, \qquad \kappa = 0.16. \tag{58}$$

The persistent reference is then updated as

$$\widehat{L}_k = \nu_k \left[\widehat{L}_{k-1} + \kappa(A_k^+ - A_k^-) \right] + (1 - \nu_k)L_k. \tag{59}$$

Activity is measured against the evolving reference. Invalid states reset that reference to the current log intensity. At an invalid-to-valid transition, $V_k = 0$ gives $A_k^+ = A_k^- = 0$, and the update retains the reference established at the immediately preceding invalid state.

Coordinate convention and aggregation. We match the active-origin coordinate convention of the observed N-Caltech101 recordings by canonicalizing each synthetic candidate response. Define

$$A(y, x) = \mathbb{I}\left[\sum_{t,p} R(t, p, y, x) > 10^{-12} \right], \tag{60}$$

and the coordinatewise active minima

$$x_0 = \min_{A(y,x)=1} x, \qquad y_0 = \min_{A(y,x)=1} y. \tag{61}$$

The complete synthetic response volume is translated by $(-x_0, -y_0)$. The translation is computed from that candidate's response activity and places its active origin in the observed-event coordinate convention.

After active-origin canonicalization, fixed 30×38 sum pooling produces an 8×8 spatial lattice. The predicted and observed responses are accumulated into eight ordinal groups as follows.

Observed-event grouping. For each nonempty recording, let $t_{\min} = \min_i t_i$ and $t_{\max} = \max_i t_i$. Timestamps are normalized within that recording:

$$\bar{t}_i = \mathrm{clip}_{[0,1]}\left(\frac{t_i - t_{\min}}{\max(t_{\max} - t_{\min}, 1)}\right), \qquad g_i^{\mathrm{obs}} = \mathrm{clip}_{[0,7]}(\lfloor 8\bar{t}_i \rfloor).$$

The denominator floor is one unit of the stored timestamp representation. Group g contains normalized times in $[g/8, (g+1)/8)$ for $g = 0, \ldots, 6$, and the final group contains $[7/8, 1]$. An event on an interior boundary belongs to the group on its right; $\bar{t}_i = 1$ is clamped to group 7. A recording with equal timestamps assigns all its events to group 0.

Predicted-response grouping. The 97 rendered states define 96 transitions indexed globally by $j = 0, \ldots, 95$. Transition j is the state update from j to $j+1$ and contributes the activity $A^+_{j+1} + A^-_{j+1}$ from Eq. 58 to group

$$g_j^{\mathrm{pred}} = \left\lfloor \frac{8(j + 1/2)}{96} \right\rfloor = \left\lfloor \frac{j}{12} \right\rfloor.$$

Each group contains 12 consecutive transitions, with the global index continuing across all three trajectory segments. With zero-based indexing, group 2 contains transitions 24–35 and group 5 contains transitions 60–71; these groups straddle the trajectory boundaries between transitions 31/32 and 63/64, respectively. The training-derived progress tables determine rendered positions along the trajectory, and the global transition index determines group membership.

Ordinal alignment and scoring. Image group g is compared with event group g at zero offset. This ordinal alignment pairs progression along the normalized event timeline with progression through the globally indexed rendering sequence. After summing polarity, the predicted response has shape

$$R_I^{\mathrm{rot}} \in \mathbb{R}_+^{8\times 8\times 8}. \tag{62}$$

The observed response has the same group and spatial dimensions. Both responses are whole-response ℓ_1 normalized using stabilizer 10^{-12}, square-root transformed, and compared by cosine.

Similarity interpretation. For positive-mass responses, let p and q be their flattened vectors normalized to unit ℓ_1 mass. With exact cosine normalization,

$$S_{\mathrm{resp}}^{\mathrm{NCal}} = \cos(\sqrt{p}, \sqrt{q}) = \sum_u \sqrt{p_u q_u} = 1 - H^2(p, q), \qquad H^2(p, q) = \tfrac{1}{2}\|\sqrt{p} - \sqrt{q}\|_2^2. \tag{63}$$

Exact cosine cancels positive global rescaling, including the scalar from stabilized ℓ_1 normalization. Zero-mass vectors and denominator clipping follow the scoring implementation. Appendix G.4 tests perturbations to response construction.

B.4 MINI N-IMAGENET ACQUISITION-INFORMED RESPONSE

mini N-ImageNet pairs displayed source images with event recordings acquired under controlled planar translation (Kim et al., 2021). The public protocol specifies the motion family; each recording's position within the motion cycle is unspecified. The primary response averages brightness-change activity from four cardinal translations into a fixed spatial–polarity representation.

Training-data raster calibration. The raster realization is calibrated exclusively on a fixed, class-balanced training subset of 1,000 paired observations: 100 classes with 10 samples per class. Parameter selection minimizes the paired training-response objective in Eq. 65. A fixed five-fold evaluation measures calibration stability on these training pairs.

We exhaustively evaluate 81 global display transforms: three base modes (stretch, aspect-preserving letterbox, and aspect-preserving center crop), isotropic multipliers $\{0.95, 1.00, 1.05\}$, and horizontal/vertical offsets in $\{-8, 0, +8\}$ pixels. Each display transform is combined with one of nine edge displacements

$$d_{\text{edge}} \in \{2, 4, 8, 12, 16, 24, 32, 48, 64\} \text{ px}, \tag{64}$$

for $81 \times 9 = 729$ calibration configurations. For each paired training sample, the image side uses a threshold-free, phase-collapsed four-direction response and the event side uses the paired polarity-split activity response. For configuration ξ, the class-balanced objective is

$$\mathcal{L}_{\text{geom}}(\xi) = \frac{1}{100} \sum_{c=1}^{100} \frac{1}{10} \sum_{i=1}^{10} \left[1 - \cos\Big(R^{\xi}_{I,c,i}, R_{E,c,i}\Big)\right]. \tag{65}$$

The selected configuration is

$$\text{stretch on } 640 \times 480, \qquad s = 1.05, \qquad \Delta_{\text{disp}} = (+8, -8) \text{ px}, \qquad d_{\text{edge}} = 12 \text{ px}. \tag{66}$$

The documented original square trajectory has a 3 mm diagonal (Kim et al., 2021), giving an edge length of $3/\sqrt{2}$ mm. Dividing the calibrated 12-pixel displacement by this edge length yields 5.656854 px/mm. In the controlled-motion comparisons, the observed recordings follow the benchmark's 6 mm motion setting. The response direction set follows the documented axis and retains the training-calibrated 12-pixel extent. These comparisons evaluate axis-conditioned responses at a common construction scale.

Phase-agnostic translational response. The frozen primary direction set is

$$\mathcal{D} = \{\rightarrow, \uparrow, \leftarrow, \downarrow\}, \tag{67}$$

with equal weight. For direction $d = (d_x, d_y)$, 65 translated states are generated along the 12-pixel edge displacement,

$$(x_k, y_k) = (x + 0.5, y + 0.5) + \frac{k}{64}\, 12(d_x, d_y), \qquad k = 0, \ldots, 64. \tag{68}$$

After the frozen display mapping $\mathcal{G}_{\text{mini}}$,

$$W_{d,k}(I) = \text{Translate}\left(\mathcal{G}_{\text{mini}}(I), \delta_{d,k}\right), \qquad L_{d,k} = \log(\max(W_{d,k}(I), \varepsilon_L)). \tag{69}$$

Consecutive log-intensity changes are

$$\Delta L_{d,k} = L_{d,k} - L_{d,k-1}, \tag{70}$$

with polarity-preserving activity

$$A^{-}_{d,k} = [-\Delta L_{d,k}]_+, \qquad A^{+}_{d,k} = [\Delta L_{d,k}]_+. \tag{71}$$

All translational variants use the same frozen nominal region $\Omega_{\text{valid}} = [64, 576) \times [64, 416)$ in (x, y) display coordinates. This region is fixed across all direction sets. Activity within this region is sum-pooled to the 8×8 response lattice. For response cell c,

$$R^{-}_{d}(c) = \sum_{k=1}^{64} \sum_{(x,y) \in c \cap \Omega_{\text{valid}}} [-\Delta L_{d,k}(x, y)]_+, \tag{72}$$

$$R^{+}_{d}(c) = \sum_{k=1}^{64} \sum_{(x,y) \in c \cap \Omega_{\text{valid}}} [\Delta L_{d,k}(x, y)]_+. \tag{73}$$

The direction-specific response is $R_d(c) = [R^{-}_{d}(c), R^{+}_{d}(c)]$ and the primary response is

$$R^{\xi}_{I}(c) = \frac{1}{4} \sum_{d \in \mathcal{D}} R_d(c), \qquad R^{\xi}_{I} \in \mathbb{R}^{2 \times 8 \times 8}_{+}. \tag{74}$$

The observed event stream is pooled over the same retained region and 8×8 lattice with separate negative/positive event-activity channels. Both responses are whole-response ℓ_1 normalized using stabilizer 10^{-12}, flattened, and compared directly by cosine.

For the controlled conditions in Appendix E, the horizontal and vertical direction sets are, respectively,

$$\mathcal{D}_H = \{+x, -x\}, \qquad \mathcal{D}_V = \{+y, -y\}, \qquad \mathcal{D}_C = \{+x, -x, +y, -y\}. \tag{75}$$

The documented axis selects the two-direction average for each controlled condition. Cardinal and axis-specific variants share the display mapping, 12-pixel displacement, 64 transitions, retained region, polarity channels, normalization, and cosine comparison.

Table 5 summarizes the two acquisition responses.

Table 5: **Frozen acquisition-response configuration and provenance.** P denotes benchmark protocol, T training-data calibration, and F fixed response construction. Mixed sources apply to the stated components.

Setting	Value / specification	Source
Shared		
Acquisition basis	Controlled sensing family documented by benchmark	P
Acquisition specification	Documented motion family, shared calibration, and fixed rendering settings	P/T/F
N-Caltech101		
Acquisition mechanism	Controlled saccadic rotation	P
Rotational realization	$(-0.5^\circ, +0.5^\circ) \to (0^\circ, -0.5^\circ)$ $\to (+0.5^\circ, +0.5^\circ) \to (-0.5^\circ, +0.5^\circ)$	P
Acquisition progress	Shared training event-rate CDF over three trajectory segments	T
Camera/raster geometry	Training-calibrated scale 160; fixed principal point $(151.5, 119.5)$	T/F
Log-intensity floor	$\varepsilon_L = 1/255$	F
Response discretization	97 states / 96 transitions; 8 ordinal groups; 12 transitions per group	F
Threshold / validity	$\kappa = 0.16$; $\tau_M = 0.999$	F
Spatial aggregation	Active-origin; 30×38 sum pooling; 8×8 output	F
Output	$R^{\mathrm{rot}} \in \mathbb{R}_+^{8\times8\times8}$; polarity collapsed	F
mini N-ImageNet		
Acquisition mechanism	Controlled planar translation	P
Calibration data	1,000 class-balanced training pairs	T
Calibration search	81 display transforms $\times$ 9 edge displacements	T
Display mapping	640×480 stretch; scale 1.05; offset $(+8, -8)$ px	T
Translation extent	12 px	T
Response discretization	4 cardinal directions; 65 states / 64 differences per direction	F
Direction / phase handling	Uniform direction average into spatial–polarity channels	F
Output	$R^{\mathrm{trans}} \in \mathbb{R}_+^{2\times8\times8}$; polarity retained	F

B.5 Static Response Representations

Simple spatial activity A. The simple activity response compares image gradient magnitude with unsigned event activity on a spatial grid. On the image side, let g_x, g_y be 3×3 Sobel responses of the grayscale image and define

$$A_I = \mathrm{Pool}_{8\times8}\left(\sqrt{g_x^2 + g_y^2 + 10^{-6}}\right). \tag{76}$$

On the event side, all events in the observation window are accumulated as unsigned activity, collapsing both time and polarity, and pooled to the same spatial lattice:

$$A_E = \mathrm{Pool}_{8\times8}\left(\sum_{n=1}^{N_E} \mathbb{I}[(x_n, y_n) = (x, y)]\right). \tag{77}$$

Its sample-dependent inputs are the grayscale image and full-recording unsigned event counts. For N-Caltech101, A follows the frozen N-Cal response-scoring convention: whole-response ℓ_1 normalization, square-root transform, and cosine. For mini N-ImageNet, it uses whole-response ℓ_1 normalization followed directly by cosine.

Static differential response G. Images use Pillow's grayscale conversion (`L`), with luminance weights $Y = 0.299I_R + 0.587I_G + 0.114I_B$. The calibrated display mapping $\mathcal{G}_{\text{mini}}$ from Appendix B.4 gives $I_{\text{disp}} = \mathcal{G}_{\text{mini}}(Y)$, using bilinear sampling and zero padding. Define

$$L_I = \log(\max(I_{\text{disp}}, 1/255)), \qquad G_I = \mathcal{P}_R([|K_x * L_I|, |K_y * L_I|]), \tag{78}$$

using the 3×3 Sobel kernels

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \qquad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \tag{79}$$

with one-pixel zero padding. For $E = \{(x_n, y_n, t_n, p_n)\}_{n=1}^{N_E}$, the full-window unsigned count map and event response are

$$H_E(x, y) = \sum_{n=1}^{N_E} \mathbb{I}[(x_n, y_n) = (x, y)], \tag{80}$$

$$G_E = \mathcal{P}_R([|K_x * H_E|, |K_y * H_E|]). \tag{81}$$

The Sobel kernels act directly on the event counts H_E. On both sides, $\mathcal{P}_R$ retains the frozen region $[64, 576) \times [64, 416)$ and sum-pools to the same 8×8 lattice. The two channels contain absolute horizontal and vertical derivatives. Responses are normalized as $\bar{G} = G/(\sum_u G_u + 10^{-12})$, flattened, and compared directly by cosine. The same G construction, with shared acquisition-response geometry and pooling, is used for Original Test and controlled-motion comparisons.

C Reproducibility and Evaluation Protocol

C.1 Datasets and Paired Retrieval Splits

Each image and event recording form a one-to-one positive pair. For a split containing N pairs, image-to-event retrieval uses all N images as queries against all N event recordings, and event-to-image retrieval reverses those roles. Each query has one paired positive and $N - 1$ other gallery observations. Throughout this appendix, *validation* denotes our internal development partition.

For N-Caltech101 (Orchard et al., 2015), images and event recordings are paired by class identity and exact filename stem. The paired pool contains 8,709 instances from 101 classes. A fixed class-stratified 70/15/15 split yields 6,090 training, 1,308 validation, and 1,311 test pairs.

For mini N-ImageNet (Kim et al., 2021), pairs are identified by released split, synset, and exact filename stem. The retained official training pool contains 129,395 pairs. Within each synset, a fixed seed-42 permutation assigns 116,455 pairs to training and 12,940 to our internal validation split. The released mini original validation split supplies *Original Test*, containing 5,000 pairs with 50 per synset. The training, internal validation, and Original Test partitions have separate paired identities and share the same class set.

Controlled motion subsets. The original training and original validation data use the same nominal acquisition regime. For the controlled response comparisons, we use the mini subsets of released Validation 3 (*Vertical*) and Validation 4 (*Horizontal*) (Kim et al., 2021). Both use 5 Hz planar camera motion with 6 mm amplitude along the specified axis. Original Test, Vertical, and Horizontal share all 5,000 image identities. Each image is paired with the event recording acquired under the corresponding condition.

Within a condition, models use the same pairs and full galleries; each checkpoint supplies its own candidates for paired response comparisons. Table 6 gives the mini partition roles.

Table 6: **mini N-ImageNet partitions and evaluation roles.** Each pair provides one query in each retrieval direction. Vertical and Horizontal share the image identities of Original Test.

Partition	Pairs	Role
Training	116,455	Initial-model training and response calibration
Internal validation	12,940	Response analyses and fusion-weight selection
Original Test	5,000	Primary retrieval, static-response comparisons, and fixed-weight fusion
Vertical	5,000	Fixed-readout comparisons at 5 Hz, 6 mm vertical motion
Horizontal	5,000	Fixed-readout comparisons at 5 Hz, 6 mm horizontal motion

C.2 Initial-Model Training and Evaluation Settings

The primary retrieval comparison uses seeds $\{1, 42, 1000\}$. For every initial model, the evaluated terminal checkpoints are at 600 epochs on N-Caltech101 and 100 epochs on mini N-ImageNet. Table 7 gives the optimizer and batch settings.

Table 7: **Initial-model training settings.** LR denotes learning rate. EventBind and CEIA report global contrastive batch sizes; PP-UOT reports physical batch size and gradient-accumulation steps. Each model uses 600 epochs on N-Caltech101 and 100 on mini N-ImageNet. For CEIA, the listed rate is the cosine schedule's base/peak; initialization and the first epoch use 5×10^{-7}.

Model / data	Optimizer	LR / peak LR	Weight decay	Batch	Epochs
EventBind / N-Cal	AdamW	10^{-6}	0.05	Global 64	600
EventBind / mini	AdamW	10^{-6}	0.05	Global 64	100
CEIA / N-Cal	AdamW	5×10^{-4}	10^{-2}	Global 128	600
CEIA / mini	AdamW	5×10^{-4}	10^{-2}	Global 128	100
PP-UOT / N-Cal	Adam	8×10^{-4}	0	128; accum. 1	600
PP-UOT / mini	Adam	8×10^{-4}	0	64; accum. 2	100

EventBind. EventBind uses CLIP ViT-L/14 initialization (Radford et al., 2021). Its RGB CLIP encoder and image-side prompt pathway remain frozen. The event backbone and event-specific modality, temporal, and learnable prompt components are optimized using AdamW (Loshchilov & Hutter, 2019) with the settings in Table 7. The event input comprises five equal-event-count frames. On N-Caltech101, the configured budget is 15,000 events per frame, for 75,000 events across the five frames. On mini N-ImageNet, it is 40,000 events per frame, for 200,000 events in total.

Event preprocessing follows the dataset-specific branches of the released EventBind implementation (Zhou et al., 2024), with matching training and evaluation settings.

CEIA. CEIA uses CLIP ViT-L/14 initialization (Xu et al., 2025; Radford et al., 2021) with a frozen RGB encoder. On the event branch, Low-Rank Adaptation (LoRA) adapters (Hu et al., 2022) use rank 16 and scaling factor 32 on the query and value projections in all 24 visual-transformer blocks. Optimized parameters include these adapters, event-encoder biases, and the event logit scale. Polarity is converted to $\{-1, +1\}$ before polarity-specific counting. The event input is a single image of polarity-specific counts accumulated across the selected events, with time collapsed. The event-count cap is 20,000 on N-Caltech101 and 70,000 on mini N-ImageNet. Training uses symmetric bidirectional contrastive classification, global batch size 128, and AdamW with a cosine learning-rate schedule. The schedule's base/peak rate is 5×10^{-4}. The optimizer is initialized at 5×10^{-7}, the first epoch uses this rate, and subsequent epochs follow the configured cosine schedule. Weight decay is 10^{-2} for the regularized parameter group; biases, normalization-gain parameters, and the logit scale use zero decay.

PP-UOT. PP-UOT is trained from scratch using Eq. 50. Both datasets use Adam (Kingma & Ba, 2015) with constant learning rate 8×10^{-4} from the first optimizer step, $\beta = (0.9, 0.999)$, optimizer $\epsilon = 10^{-8}$, and zero weight decay. N-Caltech101 uses physical batch size 128 and one accumulation

step; mini N-ImageNet uses physical batch size 64 and two accumulation steps. Training reaches 28,800 and 91,000 optimizer updates, respectively. The image/event support budgets are $(32, 40)$ and $(48, 60)$. Training proceeds to these predetermined endpoints using the paired training objective.

Response construction and study scope. The event budgets above apply to Stage I. Stage II uses the observation aggregations in Appendix B; response calibration uses training data, and construction settings are fixed with training and validation data. The primary acquisition-informed test comparison uses all three models and seeds. The follow-up static comparison uses the three EventBind and CEIA checkpoints on mini Original Test. Controlled-motion comparisons use three PP-UOT checkpoints and CEIA seed 42 under the protocol in Appendix C.5.

N-Caltech101 group reordering uses PP-UOT seed 42 on internal validation. Magnitude perturbations and shortlist-size sensitivity use this seed on each dataset's validation split. Spatial-layout controls, acquisition-group summation, and the fusion-weight sweep use all three PP-UOT validation checkpoints. Appendix C.6 defines the follow-up fixed-weight test comparison. All follow-up comparisons keep the trained model, response mappings, and candidates fixed on the previously evaluated partitions.

C.3 Metrics and Aggregation

Recall@K (R@K) is the percentage of queries whose paired observation appears among the first K results. We report $K \in \{1, 5, 10\}$, with R@1 giving top-1 accuracy. The primary summary averages the two directions:

$$\mathrm{AvgR@1} = \frac{1}{2}\left(\mathrm{R@1}_{I\to E} + \mathrm{R@1}_{E\to I}\right). \tag{82}$$

We evaluate each checkpoint and direction separately, then report arithmetic means and population standard deviations across seeds. Paired gains are computed within each checkpoint before aggregation. Single-checkpoint studies report that model's results.

Recall values are percentages and differences are percentage points (pp); all statistics use unrounded values before display. Captions identify variation across seeds, reorderings, or perturbation draws. At $K = 5$, R@5 and R@10 follow the preservation property in Appendix B.1.

Rank convention at the export cutoff. For one CEIA seed-1 E→I query on mini N-ImageNet, the stored full-gallery paired-item rank is 10 and the exported top-ten list omits that item. R@10 uses the full-gallery ranks of the original evaluator. The query lies outside Top-5, so its response-only and fused outcomes, R@1, and R@5 remain unchanged.

C.4 Paired-Query Statistical Inference

Paired 95% confidence intervals use 10,000 bootstrap resamples of paired identities (Efron, 1979). Each replicate samples identities with replacement and shares those indices across compared systems, directions, and checkpoints. Intervals are the empirical 2.5th–97.5th percentiles and describe query-level uncertainty conditional on the evaluated checkpoints; across-seed standard deviations describe variation among those checkpoints.

For model m, condition a, seed set $\mathcal{S}_m$, and evaluated directions $\mathcal{D}_{\mathrm{ret}}$, let $Y^U_{i,a,s,d} \in \{0, 1\}$ indicate top-1 correctness under rule U. For a comparison $U - V$, define

$$\begin{aligned}
\bar{\delta}^{U,V;m}_{i,a} &= \frac{1}{|\mathcal{D}_{\mathrm{ret}}|\,|\mathcal{S}_m|} \sum_{s\in\mathcal{S}_m} \sum_{d\in\mathcal{D}_{\mathrm{ret}}} \left(Y^U_{i,a,s,d} - Y^V_{i,a,s,d}\right), \\
\Delta^*_{a,m} &= \frac{100}{N} \sum_{i\in\mathcal{I}^*} \bar{\delta}^{U,V;m}_{i,a},
\end{aligned} \tag{83}$$

where N is the number of paired identities and $\mathcal{I}^*$ the resampled multiset. Bidirectional comparisons use both directions; directional comparisons use one. The primary, static-response, and fixed-fusion test comparisons use $\mathcal{S}_m = \{1, 42, 1000\}$, with $N = 1311$ for N-Caltech101 and $N = 5000$ for mini Original Test. Controlled-motion comparisons use $N = 5000$, three PP-UOT seeds, and

CEIA seed 42. Thus, each bidirectional identity-level difference averages six outcomes for three checkpoints and two for a single checkpoint.

The follow-up static-response and fixed-fusion comparisons use bootstrap seed 20270913; fusion uses the PCG64 generator. Each contrast has its own interval. Counts such as $0/10{,}000$ report resampled differences at or below zero. For the pooled PP-UOT motion comparison, each identity's Vertical and Horizontal differences are averaged before resampling, using common sampled indices. The PP-UOT $F_S - B$ increment and the CEIA controlled comparisons also use bootstrap seed 20270913.

C.5 Fixed-Readout Protocol on Controlled Motion Subsets

Vertical and Horizontal each contain the same 5,000 image identities with condition-specific event recordings (Appendix C.1). PP-UOT uses seeds $\{1, 42, 1000\}$ at epoch 100 and support budgets $(48, 60)$; CEIA uses seed 42 at epoch 100. Each checkpoint supplies its own full ranking and top-five candidates for the paired response comparisons.

Standardization and nested readouts. The standardization in Eq. 4 uses $\varepsilon_z = 10^{-8}$ and the candidate-wise quantities

$$\mu_q(X) = \frac{1}{K}\sum_{j=1}^{K} X_{q,j}, \qquad \sigma_q(X) = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(X_{q,j} - \mu_q(X))^2}, \qquad K = 5. \tag{84}$$

Let C be correspondence, A simple spatial activity, S acquisition-informed similarity, and G static differential similarity (Appendices B.4 and B.5). The background and two levels of comparison are

$$B = 0.25\, z(C) + 0.75\, z(A), \tag{85}$$

$$F_X = 0.75\, z(B) + 0.25\, z(X), \tag{86}$$

$$R_X = 0.75\, z(F_G) + 0.25\, z(X), \qquad X \in \{A, C, S, G\}. \tag{87}$$

Thus $B_G = F_G$ is the background for the second level. Each input, including every intermediate background, is standardized again within the same candidates. PP-UOT evaluates all four choices of X; CEIA evaluates S and G. The first-level contrast $F_G - F_S$ compares response choices on the same background. The second-level contrast $R_S - R_G$ compares adding acquisition similarity with reusing static evidence already present in F_G. Figure 3 summarizes these two comparisons.

Fixed weights, responses, and inference. The nested weights follow the configuration fixed on internal mini N-ImageNet validation: 0.25 weights correspondence in B and the added score at each later level. Both motion conditions and models use these coefficients. The direct correspondence–response rule in Appendix C.6 separately assigns 0.25 to correspondence and 0.75 to the response. For S, internal validation uses the cardinal response; each motion subset uses the two directions along its documented axis. Rendering, normalization, and similarity settings remain fixed. Paired inference follows Appendix C.4, using $\mathcal{S}_{\text{PP-UOT}} = \{1, 42, 1000\}$ and $\mathcal{S}_{\text{CEIA}} = \{42\}$.

C.6 Fixed-Weight Fusion Protocol on the Original Test Partitions

The test rule is Eq. 5 with correspondence weight $\lambda = 0.25$, selected in the PP-UOT validation sweep (Appendix F.2). It remains fixed across datasets, models, seeds, directions, and response choices.

Acquisition-informed fusion covers all three models on both original test partitions; static fusion covers EventBind and CEIA on mini Original Test. All use seeds $\{1, 42, 1000\}$ and the terminal checkpoints in Appendix C.2: eight settings, 24 checkpoint–response comparisons, and 18 distinct checkpoints. N-Caltech101 uses the rotational response; mini Original Test uses the cardinal response and static G.

These follow-up comparisons reuse unrounded scores from previously evaluated test partitions. Initial and response-only endpoints reproduce the earlier results. Exact ties retain initial order; top-five membership and all lower positions remain fixed. The two paired contrasts are fusion minus the corresponding response-only ranking and fusion minus initial ranking, using Appendix C.4.

D Complete Retrieval Evaluation

D.1 Seed-Level Primary Test Results

Tables 8 and 9 report all seed-level directional R@1/R@5/R@10 results for the three proposers in Table 1. Each verification row uses the same frozen acquisition-informed response and the corresponding seed-specific Stage-I candidate set. Training, pairing, and aggregation follow Appendix C. EventBind's dataset-specific preprocessing follows the official release as specified in Appendix C.2. The fixed-weight fusion results in Appendix F.3 use the same checkpoints and candidate sets.

Table 8: **Seed-level primary test retrieval on N-Caltech101.** Values are percentages. Verification uses each seed's frozen candidate set.

Method	Seed	E→I			I→E			Avg R@1
		R@1	R@5	R@10	R@1	R@5	R@10	
EventBind	1	66.438	93.898	97.407	70.862	95.423	98.474	68.650
	42	68.116	94.813	98.246	70.862	95.576	98.780	69.489
	1000	67.124	94.050	97.407	68.268	94.508	98.246	67.696
+ Verification	1	84.287	93.898	97.407	89.245	95.423	98.474	86.766
	42	85.584	94.813	98.246	89.169	95.576	98.780	87.376
	1000	84.592	94.050	97.407	88.253	94.508	98.246	86.423
CEIA	1	78.261	97.178	98.780	79.710	97.635	99.161	78.986
	42	78.566	97.254	98.703	79.024	98.093	99.542	78.795
	1000	76.506	97.178	98.703	77.879	97.178	99.390	77.193
+ Verification	1	88.177	97.178	98.780	92.296	97.635	99.161	90.236
	42	87.643	97.254	98.703	92.677	98.093	99.542	90.160
	1000	87.338	97.178	98.703	91.686	97.178	99.390	89.512
PP-UOT	1	79.786	94.355	96.949	80.244	94.966	97.483	80.015
	42	84.592	96.796	98.474	86.270	97.101	98.398	85.431
	1000	83.982	96.415	97.864	83.524	96.339	97.941	83.753
+ Verification	1	84.439	94.355	96.949	83.600	94.966	97.483	84.020
	42	85.431	96.796	98.474	85.278	97.101	98.398	85.355
	1000	85.660	96.415	97.864	85.050	96.339	97.941	85.355

On N-Caltech101, PP-UOT seed 42 changes from 85.431% to 85.355% Avg R@1; its aggregate three-seed change is $+1.843$ pp. On mini N-ImageNet, each PP-UOT checkpoint has lower initial test accuracy than its validation result and gains accuracy from verification. For every CEIA seed on mini N-ImageNet, verification improves E→I and reduces I→E accuracy. Averaging the two directional changes gives the reported bidirectional effect.

D.2 Paired-Query Confidence Intervals

Table 1 reports the six primary bidirectional gains and paired intervals under Appendix C.4. Five intervals exclude zero. CEIA mini N-ImageNet has opposing directional changes: $+4.653$ pp with interval $[3.660, 5.633]$ for E→I, and -3.367 pp with interval $[-4.500, -2.220]$ for I→E. Appendix F.4 relates these changes to correction counts.

D.3 Initial-Ranking Performance across Validation and Test

Table 10 reports the initial rankings on the internal validation and original test partitions. All values use the three prespecified seeds and terminal checkpoints. The archived three-seed results of a global dual encoder provide a second from-scratch reference; this model uses sample-level cosine retrieval. Appendix C.1 specifies the separate instance sets and gallery sizes of these partitions.

PP-UOT and the global dual encoder learn from the benchmark training pairs with random initialization, while EventBind and CEIA inherit CLIP-pretrained visual representations (Radford et al., 2021; Zhou et al., 2024; Xu et al., 2025). On mini N-ImageNet, the two from-scratch models decrease by 37.915 and 23.838 pp from internal validation to Original Test, respectively; EventBind

Table 9: **Seed-level primary test retrieval on mini N-ImageNet.** Values are percentages; R@5 and R@10 are preserved by verification.

Method	Seed	E→I			I→E			Avg R@1
		R@1	R@5	R@10	R@1	R@5	R@10	
EventBind	1	66.220	88.000	92.760	68.680	88.480	92.860	67.450
	42	66.940	87.620	92.240	67.860	88.060	92.680	67.400
	1000	65.560	87.980	92.620	67.720	88.160	92.880	66.640
+ Verification	1	79.720	88.000	92.760	74.280	88.480	92.860	77.000
	42	79.680	87.620	92.240	73.520	88.060	92.680	76.600
	1000	79.960	87.980	92.620	74.120	88.160	92.880	77.040
CEIA	1	82.520	96.460	98.000	82.860	97.040	98.480	82.690
	42	82.660	96.720	98.400	83.600	97.080	98.600	83.130
	1000	82.340	96.120	97.880	83.780	97.020	98.560	83.060
+ Verification	1	87.100	96.460	98.000	80.200	97.040	98.480	83.650
	42	87.500	96.720	98.400	79.640	97.080	98.600	83.570
	1000	86.880	96.120	97.880	80.300	97.020	98.560	83.590
PP-UOT	1	38.860	57.160	64.940	40.600	61.960	69.580	39.730
	42	41.520	61.300	68.920	42.180	62.880	70.960	41.850
	1000	41.680	61.780	69.620	41.180	61.440	70.180	41.430
+ Verification	1	52.340	57.160	64.940	52.240	61.960	69.580	52.290
	42	55.580	61.300	68.920	52.580	62.880	70.960	54.080
	1000	55.740	61.780	69.620	52.240	61.440	70.180	53.990

Table 10: **Initial-ranking validation-to-test Avg R@1.** Initial-ranking mean ± population standard deviation across three training seeds. The final column is in pp.

Dataset	Proposer	Validation	Test	Test−Validation
N-Caltech101	Global dual encoder	66.437 ± 0.848	65.561 ± 0.514	-0.876
	EventBind	68.616 ± 0.255	68.612 ± 0.732	-0.004
	CEIA	78.555 ± 0.298	78.324 ± 0.804	-0.231
	PP-UOT	83.588 ± 1.785	83.066 ± 2.264	-0.522
mini N-ImageNet	Global dual encoder	55.641 ± 2.482	31.803 ± 1.719	-23.838
	EventBind	61.980 ± 0.193	67.163 ± 0.371	+5.184
	CEIA	81.573 ± 0.442	82.960 ± 0.193	+1.387
	PP-UOT	78.918 ± 2.227	41.003 ± 0.917	-37.915

and CEIA have higher test means (Table 10). This pattern is consistent with a possible contribution of pretrained representations to cross-split robustness. The much smaller changes of the from-scratch models on N-Caltech101 also show that the pronounced degradation is specific to the mini N-ImageNet split comparison. The individual contributions of initialization, architecture, preprocessing, and partition composition remain to be separated. The fixed-weight fusion comparisons in Appendix F.3 directly evaluate each model's test candidates, including PP-UOT's 41.003% Original Test baseline.

E Response Representations and Controlled Comparisons

We expand Section 4.3 through N-Caltech101 acquisition-group interventions, Original Test response comparisons, and controlled-motion comparisons. Appendix C defines the checkpoint sets and fixed scoring rules.

E.1 N-Caltech101 Acquisition-Group Correspondence

This PP-UOT seed-42 validation comparison uses epoch 600, support budgets $(32, 40)$, and $K = 5$. It retains the eight maps examined by the group-summation control in Appendix G.1 and changes their predicted order. Nominal pairs identically indexed groups from the normalized event timeline and rendered motion sequence (Appendix B.3). Observed groups, spatial entries, response values, similarity, and candidates remain fixed. For permutation π,

$$R^{\pi}_{I,txy} = R_{I,\pi(t),x,y}, \qquad S_{\pi}(I, E) = s(R^{\pi}_{I}, R_E). \tag{88}$$

Here t indexes acquisition groups. Each control applies the same permutation to every image response. The bank contains 1,000 distinct permutations sampled uniformly from the non-identity permutations of eight groups.

Table 11: **N-Caltech101 acquisition-group reorderings.** Random statistics summarize 1,000 controls. Structured rows give the minimum–maximum range across the tested transformations. All accuracies use the PP-UOT seed-42 internal validation checkpoint.

Statistic / control	Avg R@1 (%)
Nominal ordering	88.035
Random mean $\pm$ standard deviation	84.652 ± 1.211
Random median	84.671
Random minimum	80.619
Random 97.5th percentile	86.927
Random maximum	87.576
Random results $\geq$ nominal	0 / 1,000
Reverse order	84.442
Tested cyclic shifts	82.798–85.398
Tested adjacent swaps	86.124–88.073

Nominal exceeds every sampled random result (Table 11). Its add-one tail fraction is

$$r_{\mathrm{tail}} = \frac{1 + \#\{r : \mathrm{AvgR@1}_r \geq \mathrm{AvgR@1}_{\mathrm{nominal}}\}}{1000 + 1} = \frac{1}{1001} \simeq 0.001. \tag{89}$$

This fraction describes the sampled bank at resolution $1/1001$ (Fig. 2). The strongest tested adjacent swap exceeds nominal by 0.038 pp. The combined pattern supports useful coarse group correspondence with tolerance to a local change in order.

E.2 Response Comparison on Original mini N-ImageNet Test Data

We compare S and G for EventBind and CEIA on mini Original Test using the three checkpoints and fixed-candidate protocol in Appendix C. Both responses share spatial preprocessing and sort raw response similarities. Table 12 gives per-seed bidirectional results; Tables 13 and 14 give directional means, standard deviations, and paired gains.

Table 12: **Per-seed bidirectional Original Test results.** Each model uses its epoch-100 checkpoints at seeds $\{1, 42, 1000\}$. Initial, S-only, and G-only values are Avg R@1 (%); Δ_S and Δ_G compare each verifier with the initial ranking, and all differences are in pp.

Model	Seed	Initial	S-only	G-only	Δ_S	Δ_G	$G - S$
EventBind	1	67.450	77.000	80.570	+9.550	+13.120	+3.570
	42	67.400	76.600	80.210	+9.200	+12.810	+3.610
	1000	66.640	77.040	80.580	+10.400	+13.940	+3.540
CEIA	1	82.690	83.650	88.330	+0.960	+5.640	+4.680
	42	83.130	83.570	88.210	+0.440	+5.080	+4.640
	1000	83.060	83.590	88.090	+0.530	+5.030	+4.500

Table 13: **Directional Original Test accuracy across three seeds.** Mean $\pm$ population standard deviation of R@1 (%). Avg first averages the two directions within each checkpoint.

Model	Direction	Initial	S-only	G-only
EventBind	I→E	68.087 ± 0.423	73.973 ± 0.327	81.020 ± 0.156
	E→I	66.240 ± 0.564	79.787 ± 0.124	79.887 ± 0.294
	Avg	67.163 ± 0.371	76.880 ± 0.199	80.453 ± 0.172
CEIA	I→E	83.413 ± 0.398	80.047 ± 0.290	88.567 ± 0.066
	E→I	82.507 ± 0.131	87.160 ± 0.257	87.853 ± 0.143
	Avg	82.960 ± 0.193	83.603 ± 0.034	88.210 ± 0.098

Table 14: **Directional response gains across three seeds.** Mean $\pm$ population standard deviation in pp. Standard deviations are computed from paired per-checkpoint differences.

Model	Direction	$S - \text{Initial}$	$G - \text{Initial}$	$G - S$
EventBind	I→E	$+5.887 \pm 0.364$	$+12.933 \pm 0.272$	$+7.047 \pm 0.245$
	E→I	$+13.547 \pm 0.678$	$+13.647 \pm 0.858$	$+0.100 \pm 0.188$
	Avg	$+9.717 \pm 0.504$	$+13.290 \pm 0.477$	$+3.573 \pm 0.029$
CEIA	I→E	-3.367 ± 0.537	$+5.153 \pm 0.463$	$+8.520 \pm 0.273$
	E→I	$+4.653 \pm 0.133$	$+5.347 \pm 0.100$	$+0.693 \pm 0.213$
	Avg	$+0.643 \pm 0.227$	$+5.250 \pm 0.277$	$+4.607 \pm 0.077$

Response comparison. All six model–seed bidirectional $G - S$ differences are positive. The advantage is larger in I→E (7.047/8.520 pp for EventBind/CEIA) than E→I (0.100/0.693 pp). For CEIA, G improves both directional means; S yields opposing changes. The positive $G - S$ and $G - \text{Initial}$ aggregate intervals in Table 15 support the strong-static comparison. CEIA's $S - \text{Initial}$ aggregate interval includes zero, consistent with its directional imbalance.

Table 15: **Paired uncertainty for three-seed Original Test gains.** Differences and 95% confidence intervals are in pp. The final column counts nonpositive differences in 10,000 paired identity bootstrap replicates, conditional on each model's three checkpoints.

Model	Avg contrast	Difference	95% CI	Resamples ≤ 0
EventBind	$G - S$	+3.573	[3.137, 4.017]	0 / 10,000
	$S - \text{Initial}$	+9.717	[8.783, 10.643]	0 / 10,000
	$G - \text{Initial}$	+13.290	[12.423, 14.157]	0 / 10,000
CEIA	$G - S$	+4.607	[4.130, 5.080]	0 / 10,000
	$S - \text{Initial}$	+0.643	[-0.240, 1.520]	750 / 10,000
	$G - \text{Initial}$	+5.250	[4.450, 6.057]	0 / 10,000

Appendix F.3 compares each of these responses with fixed-weight test fusion; Appendix F.2 gives the validation sweep used to choose that weight.

E.3 Controlled mini N-ImageNet Acquisition Conditions

The released Vertical and Horizontal conditions share 5,000 image identities with Original Test and use condition-specific event recordings. Appendix C.5 defines the checkpoint sets and paired candidate comparisons.

Table 16: **Frozen PP-UOT retrieval under controlled acquisition conditions.** Three-seed bidirectional means; all metrics are reported on a 0–100 scale.

Condition	Avg R@1	Avg R@5	Avg R@10	MRR
Vertical	68.813	83.823	88.117	75.640
Horizontal	44.027	62.520	69.710	52.781

Table 16 reports PP-UOT's initial retrieval; mean reciprocal rank (MRR) is the mean reciprocal paired-item rank, scaled by 100. CEIA's initial and verified endpoints appear in Table 19. The following PP-UOT comparison also includes its 12,940-pair internal validation reference. Comparisons are paired within each condition; S uses cardinal directions on validation and the documented axis on each motion subset.

E.4 Static and Repeated-Score Comparisons with PP-UOT

On internal validation, simple activity A reranks PP-UOT's candidates at 86.370% mean Avg R@1. We compare A, C, S, G on the correspondence–activity background B, then repeat the comparison after G is included, using Appendix C.5's fixed readouts. Repeated-score branches measure the gain obtainable from another use of an existing score.

At the first level, S improves on the correspondence–activity background and on the repeated A and C controls. Choosing G yields still higher accuracy in both motion conditions, with $F_S - F_G$ intervals below zero. After G is present, adding S to PP-UOT's Horizontal background gives $+0.110$ pp relative to F_G, with interval $[0.013, 0.210]$. Reusing G gives the higher mean endpoint. For $R_S - R_G$, the Vertical interval includes zero and the Horizontal interval lies below zero. These fixed-readout results make reweighting existing static evidence a strong reference for assessing an additional response score.

Paired aggregation. With the pooled-condition inference in Appendix C.4, $F_S - B$ is $+0.617$ pp with interval $[0.532, 0.702]$. Table 18 separately reports each condition's strong-static and repeated-score comparisons.

Table 17: **Complete endpoints for the fixed attribution readouts.** Three-seed PP-UOT bidirectional Avg R@1 (%). Internal validation uses 12,940 paired instances; each motion subset uses its corresponding 5,000 pairs. Comparisons are paired within each column. The readouts F_X and R_X follow Appendix C.5.

Readout	Internal validation	Vertical	Horizontal
C	78.918	68.813	44.027
B	88.536	81.723	60.787
F_A	88.402	81.717	60.987
F_C	87.709	80.333	58.477
F_S	89.147	82.347	61.397
F_G	89.458	82.513	61.580
R_A	89.194	82.333	61.580
R_C	88.539	81.293	59.650
R_S	89.503	82.590	61.690
R_G	89.701	82.630	61.773

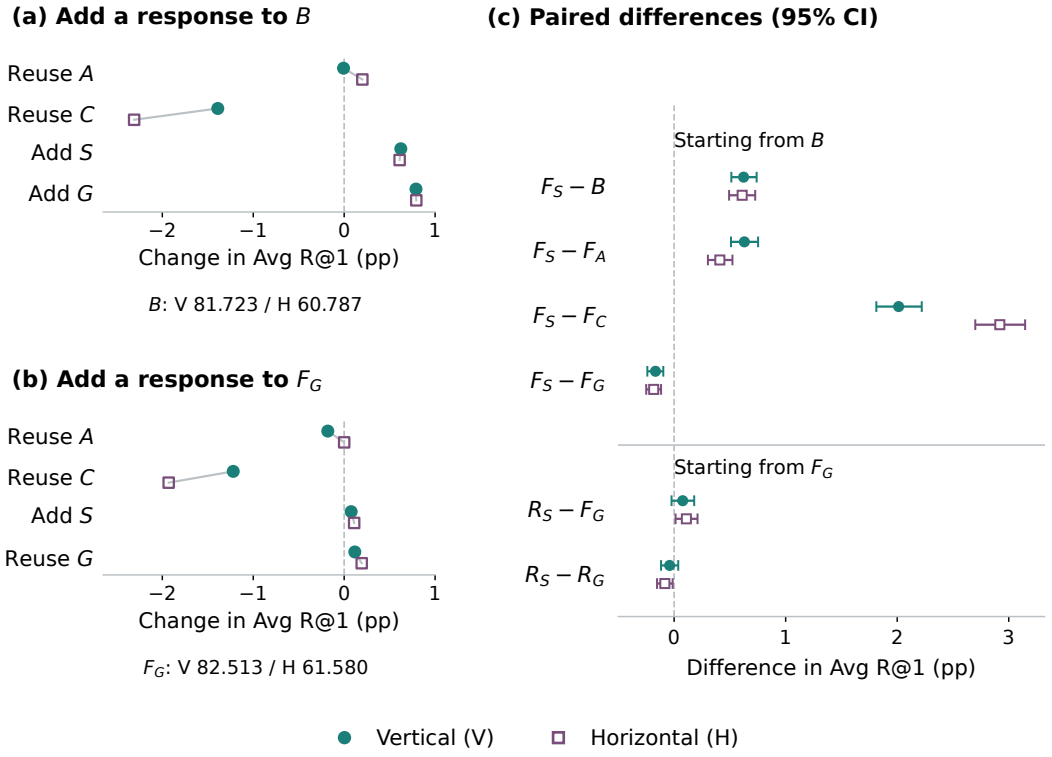


Figure 5: **PP-UOT static and repeated-score comparisons.** Panels (a,b) show within-condition changes relative to B and F_G under Appendix C.5. Connectors join the two condition-specific effects of the same readout. Panel (c) reports paired readout differences and 95% confidence intervals. Values average three PP-UOT training seeds. Complete endpoints and intervals are listed in Tables 17 and 18.

E.5 Fixed-Readout Comparisons with CEIA Candidates

CEIA uses the same B, F_X, R_X readouts with its seed-42 epoch-100 checkpoint on both motion subsets (Appendix C.5). This comparison examines the response effects on another model's candidates.

Table 18: **Paired PP-UOT fixed-readout differences.** Three-seed bidirectional changes and 95% confidence intervals in pp. Each row is the first named readout minus the second.

Comparison	Vertical	Horizontal
$F_S - B$	$+0.623$ $[0.513, 0.740]$	$+0.610$ $[0.493, 0.727]$
$F_S - F_A$	$+0.630$ $[0.510, 0.753]$	$+0.410$ $[0.303, 0.523]$
$F_S - F_C$	$+2.013$ $[1.813, 2.220]$	$+2.920$ $[2.700, 3.147]$
$F_S - F_G$	-0.167 $[-0.240, -0.097]$	-0.183 $[-0.250, -0.117]$
$R_S - F_G$	$+0.077$ $[-0.023, 0.180]$	$+0.110$ $[0.013, 0.210]$
$R_S - R_G$	-0.040 $[-0.117, 0.037]$	-0.083 $[-0.153, -0.013]$

Table 19: **Complete CEIA controlled-motion endpoints.** Bidirectional Avg R@1 (%) for seed 42. Each condition uses the same checkpoint-specific candidates for all readouts.

Condition	C	B	F_S	F_G	R_S	R_G
Vertical	82.580	93.000	93.960	94.240	94.380	**94.440**
Horizontal	79.500	92.050	92.810	93.180	92.990	**93.300**

The static first-level branch is stronger in both conditions, with positive $F_G - F_S$ intervals (Fig. 3); both have zero nonpositive resamples out of $10{,}000$. After G, adding S changes accuracy by $+0.140$ pp on Vertical and -0.190 pp on Horizontal; reusing G gives $+0.200$ and $+0.120$ pp. These descriptive endpoint differences accompany the paired $R_S - R_G$ intervals: Vertical includes zero and Horizontal favors reuse. The same qualitative pattern thus appears on three PP-UOT checkpoints and one CEIA checkpoint under the fixed readout.

E.6 Response-Axis Cross-Evaluation

Table 20 evaluates response-only reranking for the three PP-UOT checkpoints. Each column fixes the observed events and correspondence candidates; the response direction set varies across rows. The vertical and horizontal responses each average the two opposite translations along their named axis.

Table 20: **Response-only cross-evaluation on controlled conditions.** Three-seed PP-UOT bidirectional Avg R@1 (%). Vertical, horizontal, and cardinal responses use their fixed direction sets. Each column holds the observed events and initial candidate sets fixed.

Response	Vertical observations	Horizontal observations
Vertical response	78.663	58.693
Horizontal response	72.077	58.237
Phase-agnostic cardinal	77.217	59.023
Simple spatial activity	80.397	60.720

Under Vertical acquisition, the vertical response exceeds the horizontal response by 6.587 pp. Under Horizontal acquisition, the vertical response exceeds the horizontal response by 0.457 pp. These differences use unrounded evaluation outputs. Among the two single-axis responses, the vertical response has higher accuracy under both acquisition conditions.

F Combining Correspondence and Response Scores

We expand Section 4.4 with correction accounting, the validation weight sweep, complete fixed-weight test comparisons, and a CEIA case study linking fusion gains to correct-match retention.

F.1 Rescue–Damage Definitions and Exact Accounting

Let $r_{\mathrm{corr}}(q)$ and $r_{\mathrm{final}}(q)$ denote the paired item's ranks before and after verification. A query is *recoverable* when $2 \le r_{\mathrm{corr}}(q) \le K$, *rescued* when it is recoverable and $r_{\mathrm{final}}(q) = 1$, *retained*

when both ranks are one, and *damaged* when $r_{\mathrm{corr}}(q) = 1 < r_{\mathrm{final}}(q)$. The conditional rates are

$$\mathrm{RescueRate}(\%) = 100\frac{N_{\mathrm{rescued}}}{N_{\mathrm{recoverable}}}, \qquad \mathrm{DamageRate}(\%) = 100\frac{N_{\mathrm{damaged}}}{N_{\mathrm{retained}} + N_{\mathrm{damaged}}}. \quad (90)$$

Net accuracy uses the counts over all evaluated queries:

$$\Delta\mathrm{R@1}_{\mathrm{pp}} = 100\frac{N_{\mathrm{rescued}} - N_{\mathrm{damaged}}}{N_{\mathrm{queries}}}. \quad (91)$$

The identity is evaluated separately for every checkpoint and direction. Reported multi-seed changes average these per-checkpoint accuracy changes.

Single-checkpoint validation accounting. Table 21 uses PP-UOT seed 42 at epoch 600 on N-Caltech101 and epoch 100 on mini N-ImageNet, with support budgets $(32, 40)$ and $(48, 60)$. Each pooled row sums the two directions of that single checkpoint.

Table 21: **Directional validation correction counts for seed 42.** Recov. and Resc. denote recoverable and rescued queries. Rescue and damage rates use the denominators above; Net is rescued minus damaged. Pooled counts combine both directions of the seed-42 checkpoint.

Dataset	Direction	Recov.	Resc.	Rescue (%)	Retained	Damaged	Damage (%)	Net
N-Caltech101	I→E	144	113	78.472	1,032	96	8.511	+17
	E→I	151	130	86.093	1,028	85	7.637	+45
	Pooled	295	243	82.373	2,060	181	8.077	+62
mini N-ImageNet	I→E	1,466	1,324	90.314	9,628	881	8.383	+443
	E→I	1,299	1,207	92.918	10,150	576	5.370	+631
	Pooled	2,765	2,531	91.537	19,778	1,457	6.861	+1,074

On N-Caltech101, 62 net corrections over 2,616 directional queries yield $+2.370$ pp, corresponding to $85.665\% \rightarrow 88.035\%$ Avg R@1. On mini N-ImageNet, 1,074 net corrections over 25,880 directional queries yield $+4.150$ pp, corresponding to $82.052\% \rightarrow 86.202\%$. These single-checkpoint endpoints also identify the models used in the magnitude-perturbation experiment in Appendix G.4.

F.2 Full Correspondence–Response Fusion Sweep

With three PP-UOT checkpoints per internal validation split, we keep acquisition responses and candidates fixed and test correspondence weights 0, 0.25, 0.50, 0.75, and 1. The split sizes and aggregation follow Appendix C. The score follows Eq. 5:

$$S_{\mathrm{fuse},qj}(\lambda) = \lambda z_q(S_{\mathrm{corr},qj}) + (1-\lambda) z_q(S_{\mathrm{resp},qj}). \quad (92)$$

For $\lambda = \{0, 0.25, 0.50, 0.75, 1\}$, the N-Caltech101 accuracies are 87.551, 91.463, 90.048, 86.570, and 83.588% Avg R@1. The corresponding mini N-ImageNet values are 84.283, 87.196, 85.716, 82.003, and 78.918%. The highest tested point, $\lambda = 0.25$, exceeds response-only ranking by 3.912 and 2.913 pp, respectively, and exceeds both initial rankings. This weight is then held fixed in Appendix F.3.

F.3 Fixed-Weight Fusion on the Original Test Partitions

Using Appendix C.6, we compare initial, response-only, and fused rankings in all eight settings. Table 22 gives mean accuracy and across-seed variation; Table 23 gives paired gains, their population standard deviations, and the individual bootstrap intervals defined in Appendix C.4.

Across the eight settings, fusion improves three-seed mean Avg R@1 over response-only ranking by 1.597–6.026 pp and exceeds the initial ranking. All 16 bidirectional intervals lie above zero. The static responses also gain from the same frozen combination rule.

Table 22: **Fixed-weight test fusion: accuracy endpoints.** Bidirectional Avg R@1 (%), mean ± population standard deviation across three seeds. $X = S$ denotes acquisition-informed similarity and $X = G$ static differential similarity; fusion uses $0.25\,z(C) + 0.75\,z(X)$.

Model	X	Initial	X-only	Fusion
N-Caltech101				
EventBind	S	68.612 ± 0.732	86.855 ± 0.394	89.423 ± 0.694
CEIA	S	78.324 ± 0.804	89.969 ± 0.325	92.919 ± 0.307
PP-UOT	S	83.066 ± 2.264	84.910 ± 0.629	90.936 ± 1.304
mini N-ImageNet Original Test				
EventBind	S	67.163 ± 0.371	76.880 ± 0.199	80.463 ± 0.039
CEIA	S	82.960 ± 0.193	83.603 ± 0.034	89.483 ± 0.096
PP-UOT	S	41.003 ± 0.917	53.453 ± 0.823	55.050 ± 1.057
EventBind	G	67.163 ± 0.371	80.453 ± 0.172	82.773 ± 0.128
CEIA	G	82.960 ± 0.193	88.210 ± 0.098	92.010 ± 0.226

Table 23: **Fixed-weight test fusion: paired gains.** Each entry gives the mean ± population standard deviation over three seed-specific gains, followed by its individual paired 95% percentile interval. All values are in pp. Each of the 16 contrasts has zero nonpositive replicates out of 10,000.

Model	X	Fusion$-X$	Fusion$-C$
N-Caltech101			
EventBind	S	$+2.568 \pm 0.487$ $[1.716, 3.420]$	$+20.811 \pm 0.392$ $[19.374, 22.260]$
CEIA	S	$+2.949 \pm 0.018$ $[2.212, 3.699]$	$+14.594 \pm 0.497$ $[13.272, 15.967]$
PP-UOT	S	$+6.026 \pm 0.674$ $[5.034, 7.043]$	$+7.869 \pm 1.105$ $[6.852, 8.861]$
mini N-ImageNet Original Test			
EventBind	S	$+3.583 \pm 0.160$ $[3.163, 3.993]$	$+13.300 \pm 0.396$ $[12.540, 14.060]$
CEIA	S	$+5.880 \pm 0.130$ $[5.447, 6.317]$	$+6.523 \pm 0.097$ $[5.830, 7.223]$
PP-UOT	S	$+1.597 \pm 0.250$ $[1.297, 1.900]$	$+14.047 \pm 0.148$ $[13.427, 14.660]$
EventBind	G	$+2.320 \pm 0.124$ $[1.957, 2.683]$	$+15.610 \pm 0.495$ $[14.887, 16.343]$
CEIA	G	$+3.800 \pm 0.315$ $[3.423, 4.170]$	$+9.050 \pm 0.045$ $[8.417, 9.687]$

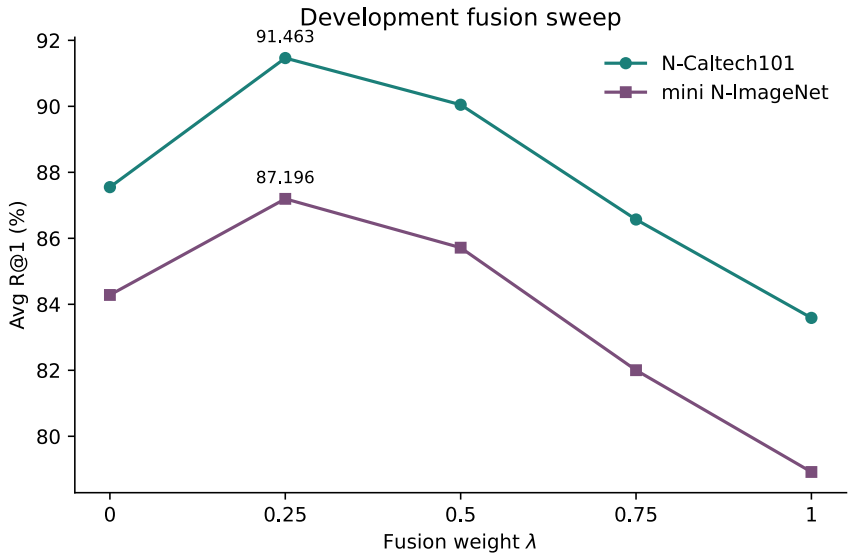


Figure 6: **Combining correspondence and response similarity.** Post-hoc internal validation results averaged over three PP-UOT seeds. $\lambda = 0$ gives response-only reranking and $\lambda = 1$ the initial correspondence ordering. The same acquisition-informed response and checkpoint-specific candidates are used across the five weights. Both datasets reach their highest evaluated accuracy at $\lambda = 0.25$.

Seed-specific and directional coverage. Tables 24–26 report each checkpoint and direction. Bidirectional accuracy for a checkpoint is the arithmetic mean of its two directional accuracies. Fusion exceeds both endpoint rankings in all 24 checkpoint–response bidirectional comparisons and all 48 checkpoint–response–direction comparisons; these summarize the same evaluated models and identities. The smallest directional gain over response-only ranking is 0.740 pp. Directional mean contrasts and their intervals appear in Table 27. R@5 and R@10 equal the corresponding initial-ranking values in Tables 8 and 9.

Table 24: **N-Caltech101: seed-specific S fusion.** Endpoint R@1 values are percentages; gains are in pp. Each row uses the same checkpoint, direction, and candidate set for all three rankings.

Model	Seed	Direction	Initial	S-only	Fusion	Fusion$-S$	Fusion$-C$
EventBind	1	I→E	70.862	89.245	91.686	+2.441	+20.824
EventBind	1	E→I	66.438	84.287	88.253	+3.966	+21.815
EventBind	42	I→E	70.862	89.169	91.228	+2.059	+20.366
EventBind	42	E→I	68.116	85.584	88.482	+2.899	+20.366
EventBind	1000	I→E	68.268	88.253	89.474	+1.220	+21.205
EventBind	1000	E→I	67.124	84.592	87.414	+2.822	+20.290
CEIA	1	I→E	79.710	92.296	94.432	+2.136	+14.722
CEIA	1	E→I	78.261	88.177	91.915	+3.738	+13.654
CEIA	42	I→E	79.024	92.677	94.966	+2.288	+15.942
CEIA	42	E→I	78.566	87.643	91.228	+3.585	+12.662
CEIA	1000	I→E	77.879	91.686	93.440	+1.754	+15.561
CEIA	1000	E→I	76.506	87.338	91.533	+4.195	+15.027
PP-UOT	1	I→E	80.244	83.600	89.550	+5.950	+9.306
PP-UOT	1	E→I	79.786	84.439	88.635	+4.195	+8.848
PP-UOT	42	I→E	86.270	85.278	92.067	+6.789	+5.797
PP-UOT	42	E→I	84.592	85.431	91.609	+6.178	+7.018
PP-UOT	1000	I→E	83.524	85.050	91.991	+6.941	+8.467
PP-UOT	1000	E→I	83.982	85.660	91.762	+6.102	+7.780

Table 25: **mini N-ImageNet Original Test: seed-specific S fusion.** Endpoint R@1 values are percentages; gains are in pp. Each row uses the same checkpoint, direction, and candidate set for all three rankings.

Model	Seed	Direction	Initial	S-only	Fusion	Fusion$-S$	Fusion$-C$
EventBind	1	I→E	68.680	74.280	79.100	+4.820	+10.420
EventBind	1	E→I	66.220	79.720	81.860	+2.140	+15.640
EventBind	42	I→E	67.860	73.520	78.820	+5.300	+10.960
EventBind	42	E→I	66.940	79.680	82.000	+2.320	+15.060
EventBind	1000	I→E	67.720	74.120	79.300	+5.180	+11.580
EventBind	1000	E→I	65.560	79.960	81.700	+1.740	+16.140
CEIA	1	I→E	82.860	80.200	87.820	+7.620	+4.960
CEIA	1	E→I	82.520	87.100	90.880	+3.780	+8.360
CEIA	42	I→E	83.600	79.640	87.820	+8.180	+4.220
CEIA	42	E→I	82.660	87.500	91.320	+3.820	+8.660
CEIA	1000	I→E	83.780	80.300	87.660	+7.360	+3.880
CEIA	1000	E→I	82.340	86.880	91.400	+4.520	+9.060
PP-UOT	1	I→E	40.600	52.240	54.060	+1.820	+13.460
PP-UOT	1	E→I	38.860	52.340	53.080	+0.740	+14.220
PP-UOT	42	I→E	42.180	52.580	55.480	+2.900	+13.300
PP-UOT	42	E→I	41.520	55.580	56.460	+0.880	+14.940
PP-UOT	1000	I→E	41.180	52.240	54.460	+2.220	+13.280
PP-UOT	1000	E→I	41.680	55.740	56.760	+1.020	+15.080

Table 26: **mini N-ImageNet Original Test: seed-specific G fusion.** Endpoint R@1 values are percentages; gains are in pp. Each row uses the same checkpoint, direction, and candidate set for all three rankings.

Model	Seed	Direction	Initial	G-only	Fusion	Fusion$-G$	Fusion$-C$
EventBind	1	I→E	68.680	81.240	83.200	+1.960	+14.520
EventBind	1	E→I	66.220	79.900	82.240	+2.340	+16.020
EventBind	42	I→E	67.860	80.900	83.060	+2.160	+15.200
EventBind	42	E→I	66.940	79.520	82.240	+2.720	+15.300
EventBind	1000	I→E	67.720	80.920	83.360	+2.440	+15.640
EventBind	1000	E→I	65.560	80.240	82.540	+2.300	+16.980
CEIA	1	I→E	82.860	88.660	91.800	+3.140	+8.940
CEIA	1	E→I	82.520	88.000	91.580	+3.580	+9.060
CEIA	42	I→E	83.600	88.520	92.160	+3.640	+8.560
CEIA	42	E→I	82.660	87.900	92.180	+4.280	+9.520
CEIA	1000	I→E	83.780	88.520	92.260	+3.740	+8.480
CEIA	1000	E→I	82.340	87.660	92.080	+4.420	+9.740

Table 27: **Directional fixed-fusion contrasts.** Three-seed mean gains and individual paired 95% intervals (pp). The N-Caltech101 EventBind I→E fusion$-S$ comparison has 2 nonpositive replicates; the other 31 directional comparisons have zero, out of 10,000 each.

Model	X	Direction	Fusion$-X$ [95% CI]	Fusion$-C$ [95% CI]
N-Caltech101				
EventBind	S	I→E	$+1.907$ $[0.941, 2.899]$	$+20.798$ $[18.967, 22.604]$
EventBind	S	E→I	$+3.229$ $[1.958, 4.424]$	$+20.824$ $[18.917, 22.807]$
CEIA	S	I→E	$+2.059$ $[1.220, 2.949]$	$+15.408$ $[13.857, 17.035]$
CEIA	S	E→I	$+3.839$ $[2.721, 4.958]$	$+13.781$ $[12.077, 15.459]$
PP-UOT	S	I→E	$+6.560$ $[5.365, 7.780]$	$+7.857$ $[6.611, 9.128]$
PP-UOT	S	E→I	$+5.492$ $[4.348, 6.662]$	$+7.882$ $[6.662, 9.077]$
mini N-ImageNet Original Test				
EventBind	S	I→E	$+5.100$ $[4.487, 5.700]$	$+10.987$ $[10.020, 11.940]$
EventBind	S	E→I	$+2.067$ $[1.573, 2.567]$	$+15.613$ $[14.693, 16.534]$
CEIA	S	I→E	$+7.720$ $[7.087, 8.360]$	$+4.353$ $[3.467, 5.253]$
CEIA	S	E→I	$+4.040$ $[3.527, 4.573]$	$+8.693$ $[7.893, 9.493]$
PP-UOT	S	I→E	$+2.313$ $[1.860, 2.767]$	$+13.347$ $[12.533, 14.147]$
PP-UOT	S	E→I	$+0.880$ $[0.520, 1.253]$	$+14.747$ $[14.013, 15.467]$
EventBind	G	I→E	$+2.187$ $[1.727, 2.653]$	$+15.120$ $[14.227, 16.013]$
EventBind	G	E→I	$+2.453$ $[1.933, 2.967]$	$+16.100$ $[15.193, 16.993]$
CEIA	G	I→E	$+3.507$ $[3.027, 3.993]$	$+8.660$ $[7.880, 9.453]$
CEIA	G	E→I	$+4.093$ $[3.587, 4.620]$	$+9.440$ $[8.653, 10.233]$

F.4 CEIA: Response Choice, Fusion, and Correct-Match Retention

Table 28 compares the initial, response-only, and fused rankings on the same CEIA mini Original Test candidates. With S, response-only ranking lowers I→E accuracy from 83.413% to 80.047%; fusion raises it to 87.767%. The gains are 7.720 pp over S-only and 4.353 pp over the initial ranking, with intervals $[7.087, 8.360]$ and $[3.467, 5.253]$. E→I also improves, from 87.160% to 91.200%. Both directions use the same response function over their respective candidate neighborhoods. The stronger G response also benefits from fusion.

Table 28: **CEIA mini N-ImageNet Original Test: response and fusion endpoints.** R@1 (%), mean ± population standard deviation across three seeds. All rules share each checkpoint's initial ranking and candidates. Average gives the bidirectional mean.

Rule	I→E	E→I	Average
Initial ranking	83.413 ± 0.398	82.507 ± 0.131	82.960 ± 0.193
S-only	80.047 ± 0.290	87.160 ± 0.257	83.603 ± 0.034
Fusion (S)	87.767 ± 0.075	91.200 ± 0.229	89.483 ± 0.096
G-only	88.567 ± 0.066	87.853 ± 0.143	88.210 ± 0.098
Fusion (G)	92.073 ± 0.198	91.947 ± 0.262	92.010 ± 0.226

Successful corrections and displaced correct matches. All four response-based rules use the same initial CEIA ranking as the reference for the accounting in Appendix F.1. Table 29 summarizes separately calculated seed-level counts and rates, with variation across seeds. Table 30 gives the integer counts and rate denominators. The original directional losses and gains recur in all three seeds.

For I→E with S, the three checkpoints together produce 1,617 rescues and 2,122 displacements under response-only ranking, compared with 1,559 rescues and 906 displacements under fusion. These counts cover 15,000 query–checkpoint outcomes from the same 5,000 paired identities. Displacements decrease by 1,216 and rescues by 58, giving a gain of $100(1216 - 58)/15000 = 7.720$ pp over S-only. Thus, the reduction in newly introduced errors outweighs the smaller reduction in successful corrections.

Table 29: **CEIA mini N-ImageNet correction accounting.** Counts are three-seed means; rates (%) and net accuracy changes (pp) are mean ± population standard deviation. Each seed evaluates 5,000 pairs per direction. Rescue and damage use the initially recoverable and correct subsets, respectively.

Rule	Rescued	Damaged	Rescue (%)	Damage (%)	Net (pp)
I→E					
S-only	539.000	707.333	79.043 ± 1.004	16.960 ± 0.425	-3.367 ± 0.537
Fusion (S)	519.667	302.000	76.216 ± 0.667	7.241 ± 0.036	$+4.353 \pm 0.451$
G-only	607.667	350.000	89.145 ± 0.540	8.392 ± 0.100	$+5.153 \pm 0.463$
Fusion (G)	581.667	148.667	85.344 ± 0.560	3.565 ± 0.108	$+8.660 \pm 0.201$
E→I					
S-only	606.000	373.333	87.032 ± 0.687	9.050 ± 0.193	$+4.653 \pm 0.133$
Fusion (S)	576.000	141.333	82.734 ± 2.050	3.426 ± 0.115	$+8.693 \pm 0.287$
G-only	616.667	349.333	88.560 ± 0.159	8.468 ± 0.155	$+5.347 \pm 0.100$
Fusion (G)	590.667	118.667	84.836 ± 1.504	2.876 ± 0.179	$+9.440 \pm 0.283$

The same count pattern appears in the other three direction–response comparisons. For I→E with G, the reduction in displacements contributes 4.027 pp and the reduction in rescues contributes -0.520 pp, for a net gain of 3.507 pp. The corresponding net gains in E→I are 4.040 pp for S and 4.093 pp for G. These fixed-candidate comparisons show benefits from both selecting the response representation and retaining learned correspondence when using that response for ranking.

Table 30: **Seed-specific CEIA correction counts on Original Test.** Each direction uses 5,000 queries per seed. Block headings give the initially correct and recoverable counts, the denominators of damage and rescue rates. Retained denotes correct matches kept first; Net equals rescued minus damaged.

I→E

Rule	Resc.	Dam.	Retained	Net
Seed 1: correct 4,143; recoverable 709				
S-only	569	702	3,441	-133
Fusion (S)	546	298	3,845	+248
G-only	633	343	3,800	+290
Fusion (G)	601	154	3,989	+447
Seed 42: correct 4,180; recoverable 674				
S-only	533	731	3,449	-198
Fusion (S)	514	303	3,877	+211
G-only	596	350	3,830	+246
Fusion (G)	574	146	4,034	+428
Seed 1000: correct 4,189; recoverable 662				
S-only	515	689	3,500	-174
Fusion (S)	499	305	3,884	+194
G-only	594	357	3,832	+237
Fusion (G)	570	146	4,043	+424

E→I

Rule	Resc.	Dam.	Retained	Net
Seed 1: correct 4,126; recoverable 697				
S-only	610	381	3,745	+229
Fusion (S)	566	148	3,978	+418
G-only	618	344	3,782	+274
Fusion (G)	582	129	3,997	+453
Seed 42: correct 4,133; recoverable 703				
S-only	605	363	3,770	+242
Fusion (S)	572	139	3,994	+433
G-only	621	359	3,774	+262
Fusion (G)	591	115	4,018	+476
Seed 1000: correct 4,117; recoverable 689				
S-only	603	376	3,741	+227
Fusion (S)	590	137	3,980	+453
G-only	611	345	3,772	+266
Fusion (G)	599	112	4,005	+487

G Selected Additional Controls

These experiments examine spatial response organization, the controlled initial ranker, candidate-specific response matching, and sensitivity to acquisition magnitude and shortlist size. Each subsection specifies its fixed model and validation setting.

G.1 Spatial Layout and Acquisition-Group Controls

We evaluate the full acquisition-informed responses and two forms of summation using fixed PP-UOT checkpoints and internal validation candidates. Spatial summation tests the value of keeping activity at separate image locations. On N-Caltech101, acquisition-group summation additionally tests the value of keeping separate activity maps along the acquisition sequence. All comparisons use seeds $\{1, 42, 1000\}$ and $K = 5$.

Summing spatial positions. For a response R_{cxy}, the control constructs

$$R_c^{\text{sum}} = \sum_{x,y} R_{cxy}, \tag{93}$$

on both the image and event sides. The channel c indexes the eight ordered acquisition groups for N-Caltech101 and the two polarity channels for mini N-ImageNet. The control retains eight global totals on N-Caltech101 and two on mini N-ImageNet, reducing the respective response dimensions from $8\times8\times8$ to 8 and from $2\times8\times8$ to 2. Initial rankings, top-five membership, and dataset-specific whole-response normalization and similarity rules are fixed across the comparison.

Combining acquisition maps while retaining spatial layout. For N-Caltech101, we also sum the eight acquisition groups at each spatial location on both sides:

$$R_{xy}^{\text{spatial}} = \sum_{g=0}^{7} R_{gxy}.$$

The input R_{gxy} is the raw nonnegative response after polarity aggregation. The operation produces one 8×8 spatial map; whole-response ℓ_1 normalization and the elementwise square root are then applied to this map, followed by the same cosine comparison as the full response. The sum therefore precedes both normalization and the square-root transform. The original eight-map response and this single-map control use the same spatial grid, recorded response values, and candidate sets.

Table 31: **Spatial and acquisition-group response controls.** Three-seed bidirectional internal validation Avg R@1 (%) using PP-UOT candidates. Acquisition-group summation keeps the spatial grid; spatial summation keeps only the acquisition-group or polarity totals.

Response	Dimensions	Avg R@1 (%)
N-Caltech101		
Full response	$8 \times 8 \times 8$	87.551
Acquisition groups summed	8×8	86.149
Spatial positions summed	8	31.333
mini N-ImageNet		
Full response	$2 \times 8 \times 8$	84.283
Spatial positions summed	2	18.077

The spatially resolved responses have higher accuracy than their compressed per-channel totals on both benchmarks. On N-Caltech101, the full eight-map response exceeds the single spatial map by 1.402 pp in three-run mean accuracy (87.551% versus 86.149%). Table 32 reports the paired seed-level comparison. It quantifies the value of retaining acquisition groups within this response construction. The ordering intervention in Appendix E.1 separately preserves all response maps and changes their correspondence between the image and event sides.

G.2 Compact PP-UOT Proposer Comparison

The four variants in Table 33 are independently trained with seed 42 under matched dataset-specific budgets and evaluated on internal validation using their initial rankings. The global dual encoder uses sample-level cosine retrieval. PointPool computes a mass-weighted average of selected local features and compares the resulting vectors by cosine. Balanced optimal transport (OT) uses the PP-UOT supports and pairwise cost with exact transport marginals; PP-UOT uses relaxed marginal penalties.

Table 32: **N-Caltech101 acquisition-group summation by seed.** Internal-validation R@1 (%) from epoch-600 PP-UOT checkpoints. Full retains eight activity maps; Summed combines them at each spatial location before normalization. Δ is Full minus Summed (pp). Average is the two-direction mean for that checkpoint.

Seed	Direction	Initial	Full	Summed	Δ
1	I→E	81.575	86.544	85.780	+0.765
1	E→I	81.040	87.538	86.086	+1.453
1	Average	81.307	87.041	85.933	+1.109
42	I→E	86.239	87.538	86.621	+0.917
42	E→I	85.092	88.532	86.621	+1.911
42	Average	85.665	88.035	86.621	+1.414
1000	I→E	83.945	87.232	85.703	+1.529
1000	E→I	83.639	87.920	86.086	+1.835
1000	Average	83.792	87.576	85.894	+1.682

Table 33: **Controlled Stage-I mechanism comparison.** Seed-42 internal validation Avg R@1 (%) for independently trained initial-ranker variants under matched dataset-specific training budgets.

Variant	N-Caltech101	mini N-ImageNet
Global Dual-Encoder	65.558	58.814
Mass-Weighted PointPool	60.474	54.590
Balanced OT	82.301	72.732
PP-UOT	**85.665**	**82.052**

The explicit local-correspondence variants achieve higher accuracy than pooled local features in this comparison. PP-UOT also improves on balanced transport. These results characterize the initial-ranker choices under the specified training budgets.

G.3 Arbitrary Reranking and Mismatched Responses

We use PP-UOT seed 42 on internal validation, with epoch 600 for N-Caltech101 and epoch 100 for mini N-ImageNet. Appendix F.1 reports these checkpoints' initial and verified results. Across 100 trials with randomization seeds 0–99, we preserve top-five membership and randomly reorder the candidates. Table 34 reports the resulting accuracy distribution.

Table 34: **Random within-Top-5 reranking.** Mean $\pm$ standard deviation, central empirical 95% range, and best Avg R@1 (%) over 100 random reranking trials. The range contains the central 95% of trial-level accuracies.

Dataset	Avg R@1	Empirical 95% range	Best
N-Caltech101	19.387 ± 1.085	17.640–21.607	21.942
mini N-ImageNet	18.535 ± 0.301	17.983–19.132	19.208

Deterministic cyclic response mismatch. For the same mini N-ImageNet checkpoint, sort the $N = 12{,}940$ validation identities lexicographically by sample identifier, giving $i_0, \ldots, i_{N-1}$. Define the cyclic successor map

$$\tau(i_j) = i_{(j+1) \bmod N}, \qquad \widetilde{R}_I(i_j) = R_I(\tau(i_j)) .$$

Every identity is assigned the image response of its successor, and the last identity is assigned the first identity's response. The mapping is deterministic and has zero fixed points. Both retrieval directions use this same forward image-response map:

$$v_r^{\mathrm{mis}}(q) = \begin{cases} s\Big(\widetilde{R}_I(i_r), R_E(i_q)\Big), & \text{E→I}, \\ s\Big(\widetilde{R}_I(i_q), R_E(i_r)\Big), & \text{I→E}. \end{cases}$$

Thus, E→I uses the successor response of each image candidate, and I→E uses the successor response of the image query. The initial rankings, candidate identities, positive-pair labels, and observed event responses remain fixed. This single cyclic-mismatch evaluation gives 18.972% I→E and 20.085% E→I R@1, or 19.529% Avg R@1. The random-order trials and the cyclic reassignment both yield low accuracy for their stated fixed-checkpoint comparisons. Appendix E compares structured response alternatives with their candidate associations preserved.

G.4 Query-Specific Acquisition-Magnitude Perturbation

This validation experiment uses PP-UOT seed 42 on each dataset. N-Caltech101 uses epoch 600 and support budgets $(32, 40)$; mini N-ImageNet uses epoch 100 and support budgets $(48, 60)$. Initial and nominal verified Avg R@1 are 85.665% and 88.035% on N-Caltech101, and 82.052% and 86.202% on mini N-ImageNet.

For scalar perturbations, sample $\alpha_q \sim \mathcal{U}(1-\delta, 1+\delta)$ independently across queries and multiply the nominal acquisition magnitude by α_q. All frozen candidates of the same query share that multiplier. The mini axis-specific variant samples independent $\alpha_{x,q}, \alpha_{y,q}$ from the same distribution. The observed responses, Stage-I rankings, and candidate memberships are unchanged.

The diagnostic evaluates $\delta \in \{0, 0.1, 0.2, 0.3\}$ with 20 fixed perturbation draws for each nonzero level, generated from master seed 20270913. For a fixed checkpoint, the retained-gain percentage is

$$100 \frac{\overline{\mathrm{AvgR@1}}_{\mathrm{perturbed}} - \mathrm{AvgR@1}_{\mathrm{Stage\,I}}}{\mathrm{AvgR@1}_{\mathrm{nominal}} - \mathrm{AvgR@1}_{\mathrm{Stage\,I}}}. \tag{94}$$

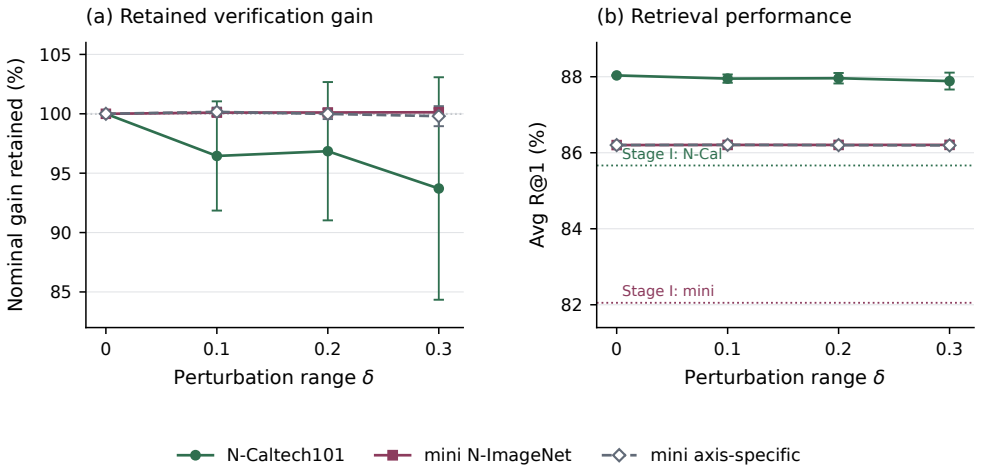


Figure 7: **Query-specific acquisition-magnitude robustness.** Panel (a) reports retained nominal verification gain; panel (b) reports Avg R@1, with dotted Stage-I baselines. At each nonzero perturbation level, points are means and error bars are ± 1 standard deviation over 20 fixed perturbation draws for a fixed trained checkpoint. The retention SD is obtained by scaling the Avg R@1 SD by the fixed nominal verification gain. Candidate sets and observed responses are fixed.

At the largest tested range, $\delta = 0.3$, N-Caltech101 reaches $87.886 \pm 0.222\%$ Avg R@1 over the 20 draws and retains 93.71% of its nominal verification gain. The mini N-ImageNet scalar condition reaches $86.207 \pm 0.013\%$ with 100.13% retained gain, while the axis-specific condition reaches $86.194 \pm 0.035\%$ with 99.80% retained.

The corresponding rescue/damage rates are $81.492/8.135\%$, $91.573/6.859\%$, and $91.618/6.882\%$, respectively. The three conditions retain most or all of the nominal verification gain at the largest evaluated perturbation range. This experiment measures sensitivity to query-specific magnitude variation within the fixed response family and calibrated spatial geometry.

G.5 VERIFICATION SHORTLIST-SIZE SENSITIVITY

This internal-validation experiment uses PP-UOT seed 42 at 600 epochs on N-Caltech101 and 100 epochs on mini N-ImageNet, with support budgets $(32, 40)$ and $(48, 60)$, respectively. The acquisition-informed response and full-gallery correspondence ranking remain fixed. For each $K \in \{1, 3, 5, 10\}$, response similarity reranks the first K candidates. The $K = 1$ condition preserves the initial prediction; the primary retrieval protocol uses the common five-candidate budget across models and datasets.

Table 35: **Shortlist-size sensitivity on internal validation.** Bidirectional Avg R@1 (%) from one seed-42 PP-UOT checkpoint per benchmark, using the fixed acquisition-informed response.

Candidates K	N-Caltech101	mini N-ImageNet
1 (initial)	85.665	82.052
3	**89.679**	**86.240**
5	88.035	86.202
10	85.894	85.062

The highest evaluated accuracy occurs at $K = 3$ on both benchmarks. Increasing K from 3 to 10 raises the number of recoverable directional queries from 246 to 336 on N-Caltech101 and from 2,184 to 3,384 on mini N-ImageNet. Over the same range, net corrections decrease from 105 to 6 and from 1,084 to 779, respectively. Candidate coverage and final top-1 accuracy therefore respond differently to shortlist expansion in this fixed-response comparison.