TYPED EVIDENCE-CONFLICT RERANKING AT VIDEO-LANGUAGE HANDOFFS

Anonymous authors
Paper under double-blind review

ABSTRACT

A reranker's full-set improvement is the prevalence-weighted sum of its gains on different query populations, but standard video-retrieval reporting leaves that accounting implicit at the evidence handoff. We introduce the Fixed-Pool Answer-Conflict Audit, which identifies semantically close candidates that disagree with answer evidence inside immutable within-source top-10 pools, and an additive decomposition that separates audited and complementary gains. The operating point is fixed before final-evaluation scoring; blinded adjudication measures proposal precision at 83.0% on TVQA and 78.5% on Ego4D-NLQ, while paired reranker gaps remain positive over two orthogonal five-point threshold sweeps. Typed Evidence Conflict Reranker (TECR) aligns Entity, Region, Utterance, Action, and Temporal evidence, then factors five localized conflicts into applicability, answer relevance, and reliability. Against a BERT-large cross-encoder, TECR improves full-set nDCG@10 by 0.8 and 1.2 points but audited nDCG@10 by 4.2 and 4.7 points; populations comprising 12.7% and 13.4% of the queries contribute 66.6% and 52.5% of those mean gains. Cross-Encoder + OpSup, Type-Agnostic Conflict, and TECR each exhibit a 3.2×–5.75× audited-to-full conditional gain ratio under the common cross-encoder reference. TECR further retains 1.9- and 1.7-point audited advantages over a supervision- and capacity-matched type-agnostic control, and Charades-STA extends both the gain decomposition and typed-structure advantage to a third corpus. These results turn the delivered evidence state into a measurable model-comparison object and show that typed alignment improves the handoffs carrying a disproportionate share of reranking utility.

1 INTRODUCTION

Video reranking acts at an evidence handoff: among already retrieved moments, it decides which candidate reaches a question answerer or localization consumer first. A full-set nDCG improvement alone cannot identify the states changed by that decision. For any fixed query population, the mean gain is exactly the prevalence-weighted sum of the gain on a designated subset and its complement. Making this identity useful therefore requires a subset defined from the candidate state and official evidence before the compared rerankers are scored.

The Fixed-Pool Answer-Conflict Audit supplies that object without changing candidate availability. One frozen ViT-L/14 retriever gives every method the same within-source top-10 list. A query is tagged when the list contains a dense near match separated from official support; TVQA additionally requires a frozen option-evidence rule to favor a non-gold answer, whereas the localization corpora instantiate the protocol as support conflict. The thresholds and all tie breaks are fixed before final-evaluation scoring, and neither audit membership nor audited metrics enter model selection. Blinded adjudication estimates 83.0% precision on TVQA and 78.5% on Ego4D-NLQ; both five-point sensitivity sweeps preserve positive paired gaps over nine unique threshold pairs.

This design exposes a stable accounting pattern rather than a favorable single comparison. With the vanilla cross-encoder as a common reference, Cross-Encoder + OpSup, Type-Agnostic Conflict, and TECR all improve audited queries 3.2×–5.75× more per query than the complete set across TVQA and Ego4D-NLQ. For TECR, 1,934 TVQA queries (12.7%) contribute 66.6% of its 0.8-point mean gain, and 687 Ego4D-NLQ queries (13.4%) contribute 52.5% of its 1.2-point gain; complement gains are 0.31 and 0.66 nDCG@10. Answer-conflict gain concentration (ACGC) names the pair-specific

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.

conditional ratio, while the additive share states how much of the full-set mean comes from the audited population.

Typed Evidence Conflict Reranker (TECR) tests whether the audited handoff admits a targeted mechanism. Frozen LLaVA-1.6-7B and GroundingDINO descriptors populate local Entity, Region, Utterance, Action, and Temporal nodes (Liu et al., 2024a;b). Type-constrained alignment produces one shared query–candidate representation, and five operator heads factor entity, attribute, temporal, region, and utterance/audio conflicts into applicability, answer relevance, and reliability. Against a supervision- and capacity-matched type-agnostic control, TECR retains 1.9 and 1.7 audited nDCG points with paired intervals excluding zero; component removals, adjudicated families, controlled edits, five seeds, and Charades-STA provide complementary tests of the combined typed-structure intervention.

The experiments are organized around three falsifiable questions. Does a reranker-invariant rule fixed before final-evaluation scoring identify genuine evidence conflicts across nearby operating points? Does gain concentration persist under a common reference rather than depend on one selected comparator? Does typed alignment retain an advantage when descriptor access, supervision, optimization, and capacity are matched? Together, the audit, additive accounting, and TECR answer these questions within a shared candidate contract while preserving the standard full-set metric.

2 Related Work

Training-time difficult examples. Counterfactual augmentation and hard-negative mining improve learning by changing examples or negative selection (Kaushik et al., 2020; Zmigrod et al., 2019; Xiong et al., 2021). Our evaluation begins after those choices: a trained retriever and its realized candidate supply are frozen, and the measured object is the ordering of one shared handoff. Synthetic operators train TECR, but they do not define membership in the audit.

Behavioral and subgroup evaluation. Contrast sets, behavioral test suites, and worst-group evaluation test specified behaviors or groups (Gardner et al., 2020; Ribeiro et al., 2020; Sagawa et al., 2020). Here the unit is a naturally realized retrieval state, and membership jointly requires semantic confusability, official-support exclusion, and a dataset-specific conflict test inside one immutable within-video list. The additive decomposition further ties the conditional effect back to the reported full-set mean; ACGC summarizes the relative per-query effect for a named method–reference pair.

Temporal grounding and negative-aware retrieval. QD-DETR learns query-conditioned video saliency using irrelevant video–query pairs, SHINE constructs hierarchical hard-negative queries for compositional grounding, and BM-DETR uses within-video negative queries to model background (Moon et al., 2023; Cheng et al., 2024; Jung et al., 2025). Negative-aware VMR instead asks whether a queried moment is absent and adds a rejection decision (Flanagan et al., 2025). These methods improve proposal or grounding behavior. Our object is complementary and more localized: after one candidate list has already been realized, the audit identifies close support conflicts within that list, while TECR compares typed evidence among those unchanged candidates.

Fixed-pool reranking and structured matching. Cross-encoders, poly-encoders, and ColBERT-style late interaction rescore shared candidates effectively (Nogueira & Cho, 2019; Humeau et al., 2020; Khattab & Zaharia, 2020; Santhanam et al., 2022). The scalar and late-interaction baselines evaluated here do not preserve an explicit type for the evidence mismatch behind a score. Scene graphs and grounded multimodal models represent local structure (Johnson et al., 2015; Krishna et al., 2017; Yang et al., 2018; Kamath et al., 2021), while corpus graph retrieval constructs a graph search substrate (Edge et al., 2024; He et al., 2024). TECR instead leaves dense retrieval unchanged, performs typed local comparison only on the realized top-10 handoff, and is tested against a type-agnostic model with the same descriptors, heads, losses, examples, and parameter budget. The new contribution is therefore the combination of a reranker-invariant handoff object, exact gain accounting on that object, and a matched test of typed conflict resolution.

3 FIXED-POOL ANSWER-CONFLICT AUDIT

For query q, a frozen retriever R_0 produces the ordered pool $C_q = \mathrm{TopK}_{m \in \mathcal{M}_{z(q)}} R_0(q, m)$ with $K = 10$, where $\mathcal{M}_{z(q)}$ contains only shards from source episode or video $z(q)$. TVQA candidates are its benchmark timestamped subtitle–video shards; Ego4D-NLQ and Charades-STA candidates are the frozen retriever's temporal proposals, so the audit adds no sliding-window duration or stride. We use the frozen CLIP ViT-L/14 checkpoint (Radford et al., 2021); $\phi(q)$ and $\phi(d)$ are its normalized query and candidate embeddings. Let $P_q^+ \subset C_q$ contain candidates with positive overlap with any official support interval, and let the dataset-specific conflict test $I_q(d)$ be defined below. The audit indicator is

$$A_{\tau_s,\tau_o}(q) = \mathbb{1}[P_q^+ \neq \varnothing]\,\mathbb{1}\big[\,\exists d \in C_q \setminus P_q^+ : z(d) = z(q),\ \cos(\phi(q), \phi(d)) \geq \tau_s,\ \max_{g \in P_q^+} \mathrm{IoU}(d, g) < \tau_o,\ I_q(d) = 1\,\big]. \tag{1}$$

For TVQA question text x_q, options $o_{q1}, \ldots, o_{q5}$, and gold option y_q, define

$$e_j(q, d) = \cos(\phi([x_q; o_{qj}]), \phi(d))\,, \quad \widehat{y}(q, d) = \arg\max_{j \in \{1,\ldots,5\}} e_j(q, d), \quad I_q(d) = \mathbb{1}[\widehat{y}(q, d) \neq y_q]. \tag{2}$$

The smallest option index resolves an exact score tie. This frozen scoring rule, including the option serialization, is fixed with (τ_s, τ_o) and does not use any evaluated reranker. For Ego4D-NLQ and Charades-STA, $I_q(d) \equiv 1$: the semantic-proximity and official-support exclusion terms in Equation 1 define a *support-conflict proposal*. The audit retains its answer-conflict name for the joint protocol motivated by TVQA, while all localization claims use support-conflict terminology.

Each query contributes once even when several candidates qualify. For adjudication and examples, its canonical candidate $d^\star$ is the qualifying candidate with highest frozen similarity, with base-retriever rank breaking ties. Queries with $P_q^+ = \varnothing$ cannot enter the audit. The IoU term measures separation between the distractor and retrieved support-bearing candidates, rather than duplicating the positive-overlap definition. Using only source-disjoint train-dev pools, we fixed $\tau_o = 0.30$ to the localization positivity boundary and $\tau_s = 0.85$ as the prespecified midpoint of the $[0.80, 0.90]$ semantic-near sensitivity range before computing any final-evaluation reranker outcome. The rule, option serialization, and tie breaks were then locked. The two orthogonal five-point sensitivity sweeps vary one threshold at a time; because they share their center, they contain nine unique parameter pairs. At this fixed operating point, the audit tags 1,934 of 15,253 TVQA queries and 687 of 5,127 Ego4D-NLQ queries. The rule proposes a population for evaluation, and the blinded study below estimates its semantic precision.

Let $\mathcal{Q}_A = \{q \in \mathcal{Q} : A_{\tau_s,\tau_o}(q) = 1\}$, let $\mu(M, S)$ denote the macro-average of per-query metric μ for reranker M on query set S, and let B be a named reference reranker. We quantify where an improvement occurs by

$$\begin{aligned}
\Delta^{\mu}_{\mathrm{all}}(M, B) &= \mu(M, \mathcal{Q}) - \mu(B, \mathcal{Q}),\\
\Delta^{\mu}_{\mathrm{AC}}(M, B) &= \mu(M, \mathcal{Q}_A) - \mu(B, \mathcal{Q}_A),\\
\Delta^{\mu}_{\overline{\mathrm{AC}}}(M, B) &= \mu(M, \mathcal{Q} \setminus \mathcal{Q}_A) - \mu(B, \mathcal{Q} \setminus \mathcal{Q}_A),\\
\Gamma^{\mu}_{\mathrm{AC}}(M, B) &= \Delta^{\mu}_{\mathrm{AC}}(M, B) \,/\, \Delta^{\mu}_{\mathrm{all}}(M, B),\\
\rho^{\mu}_{\mathrm{AC}}(M, B) &= p_A \Delta^{\mu}_{\mathrm{AC}}(M, B) \,/\, \Delta^{\mu}_{\mathrm{all}}(M, B), \quad p_A = |\mathcal{Q}_A|/|\mathcal{Q}|.
\end{aligned} \tag{3}$$

ACGC is the pair-specific descriptive ratio Γ_{AC}: a ratio of conditional mean gains, not the share of total improvement. The additive identity $\Delta_{\mathrm{all}} = p_A \Delta_{\mathrm{AC}} + (1 - p_A)\Delta_{\overline{\mathrm{AC}}}$ makes ρ_{AC} that share. We report the ratio only for positive nonzero full-set gains and always place it beside p_A, both absolute deltas, the complement gain, and ρ_{AC}; inference rests on the absolute paired differences rather than division by a small denominator.

The tagged samples are substantially more conflict-bearing than the boundary controls: TVQA's boundary precision is 52.0% (Wilson interval 42.3–61.5) and Ego4D-NLQ's is 47.0% (37.5–56.7). The other/compositional labels also measure conflicts outside the five modeled families. Figure 1 and Appendix A show that the absolute audited gap remains positive as either threshold moves. Together with fixing the center before final-evaluation scoring, this establishes robustness to the reported operating-point neighborhood.

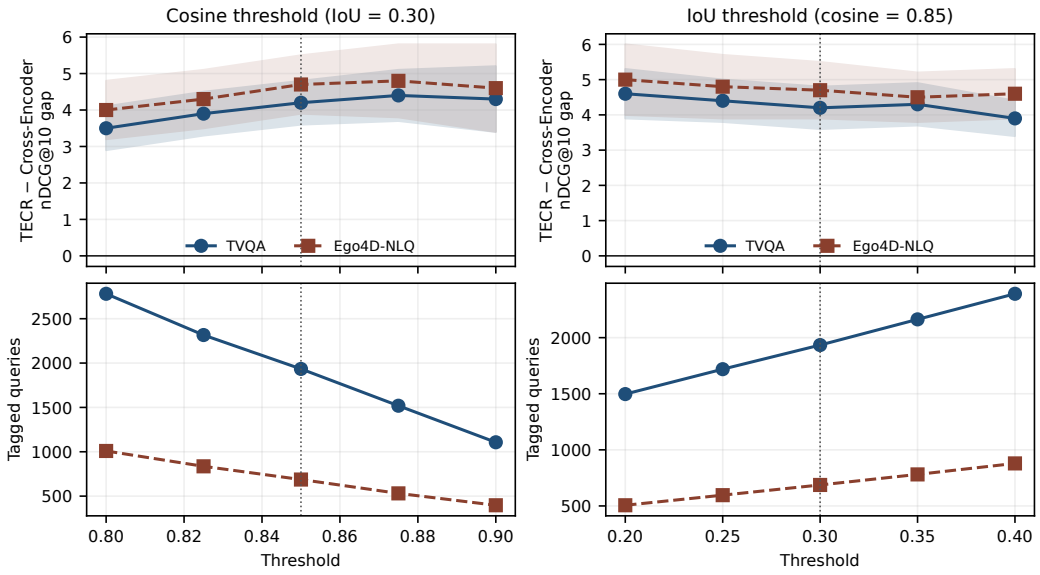


Figure 1: Audit stability on the same frozen top-10 lists. Across cosine and IoU sweeps, population size changes as intended while the TECR-minus-cross-encoder nDCG@10 gap remains positive; shaded bands are paired-query bootstrap 95% intervals from 10000 resamples. Center markers denote the fixed-before-evaluation operating point $(0.85, 0.30)$.

Table 1: Blinded audit validity. Two annotators label independently before adjudication. Wilson 95% intervals accompany exact counts; coverage counts cases explained by one of the five operator families. Boundary controls sample queries adjacent to the decision rule.

Dataset	Tagged precision	95% CI	κ	Boundary / operator coverage
TVQA	166/200 (83.0%)	77.2–87.6	0.78	52/100; 136/166
Ego4D-NLQ	157/200 (78.5%)	72.3–83.6	0.74	47/100; 119/157

4 Typed Evidence Conflict Reranker

4.1 Typed substrate and one-step conflict set

Each candidate graph $\mathcal{G}_i = (V_i, E_i)$ contains Entity, Region, Utterance, Action, and Temporal nodes. Frozen LLaVA-1.6-7B descriptors provide entity and attribute labels, GroundingDINO provides grounded regions, transcripts provide utterance spans, and timestamps provide temporal relations. Candidate graphs are constructed offline; query graphs use the same schema and leave unavailable modalities empty.

The operational hypothesis space contains exactly the evaluated one-step conflicts:

$$\mathcal{H}(\mathcal{G}_q, \mathcal{G}_i) = \{o(\mathcal{G}_i) : o \in \mathcal{O},\ d_{\text{edit}}(\mathcal{G}_i, o(\mathcal{G}_i)) = 1\},$$
$$\mathcal{O} = \{\texttt{entity_sub}, \texttt{attr_flip}, \texttt{temp_swap}, \texttt{region_rem}, \texttt{audio_drop}\}. \tag{4}$$

This single-step set supplies the operator labels used in pretraining and the conflict hypotheses scored at inference.

4.2 Shared alignment, conflict heads, and score

Nodes align only within type by one-to-one maximum-weight frozen-descriptor matching; unmatched nodes pair with a learned NULL token. A two-layer, 256-dimensional, four-head cross-attention module processes aligned pairs, type embeddings, relation tags, timestamp offsets, and missing-node

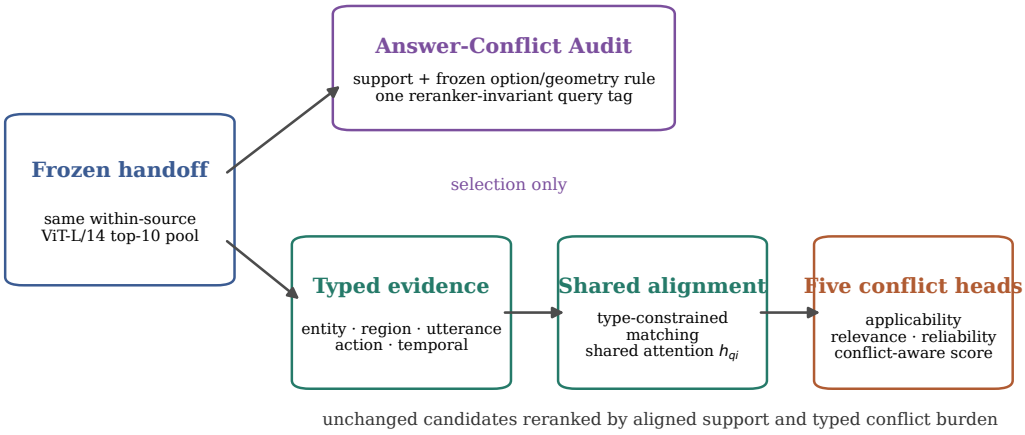


Figure 2: The audit and reranker share a frozen handoff but remain method-independent components. The audit tags a query from official support and pool geometry; TECR aligns typed query–candidate evidence once, evaluates five conflict heads on the shared representation, and reranks the unchanged candidates.

indicators once. Its pooled representation h_{qi} is shared by all operator heads. Each operator-specific head maps h_{qi} to three sigmoid outputs: applicability a_o, answer relevance r_o, and reliability c_o. All lie in $[0, 1]$ with one direction: larger values indicate stronger evidence that o explains an evidence-changing conflict. Their normalized burden is

$$B(q, m_i) = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} a_o(h_{qi}) r_o(h_{qi}) c_o(h_{qi}) \in [0, 1]. \tag{5}$$

The reliability gate c_o is learned, not interpreted as a calibrated probability. A separate aligned-support head $U(q, m_i) = \sigma(w_u^\top h_{qi})$ measures positive compatibility. After min–max normalizing base similarities within C_q to $\widetilde{s}_{\text{emb}} \in [0, 1]$, the final higher-is-better score is

$$s(q, m_i) = (1 - \alpha - \beta)\widetilde{s}_{\text{emb}}(q, m_i) + \alpha U(q, m_i) + \beta[1 - B(q, m_i)], \quad \alpha, \beta \geq 0,\ \alpha + \beta \leq 1. \tag{6}$$

We select (α, β) on a fixed source-disjoint development partition carved from the corresponding training data by full-distribution nDCG@10. Evaluation queries, audit membership, and audit metrics never enter weight or checkpoint selection. When conflict burdens are similar across the pool, the $1 - B$ term does not change candidate order and the decision is governed by base similarity and aligned support.

Synthetic one-step perturbations pretrain the three operator targets. Donors come only from TVQA training episodes; TVQA validation/test episodes, every Ego4D and Charades video, and all real evaluation shards are excluded. The loss $\mathcal{L}_{\text{op}} + 0.5\mathcal{L}_{\text{answer}} + \mathcal{L}_{\text{rank}}$ supervises the applied operator, whether its edited element overlaps query-anchored gold evidence, and the gold-versus-perturbed ranking margin; reliability receives the correctness gate from the corresponding ranking residual. Cross-Encoder + OpSup receives the same real examples, perturbed pairs, operator labels, evidence-change targets, optimizer schedule, and updates. The Type-Agnostic Conflict Transformer receives the same descriptor bundle, parameter budget, heads, losses, and examples while jointly removing node types, relation tags, and constrained alignment. It therefore tests the combined typed-structure bundle as one matched intervention.

If n_t is the node count of type t, $n = \sum_t n_t$, width $d = 256$, and layer count $L = 2$, one candidate costs $O(\sum_t n_t^3)$ for exact within-type matching, $O(L(n^2 d + n d^2))$ for shared attention, and $O(|\mathcal{O}|d)$ for the five heads. Streaming candidates requires $O(n^2 + nd)$ working memory. For $K = 10$, attention runs ten times and the shared representations feed 50 inexpensive scalar head evaluations; the implementation does not repeat cross-attention per operator.

5 EXPERIMENTS

5.1 PROTOCOL AND COMPARISON BOUNDARY

TVQA (Lei et al., 2018) contributes 15,253 test queries, Ego4D-NLQ (Grauman et al., 2022) contributes 5,127 validation queries, SynNearMiss contains 8,000 labeled perturbations, and Charades-STA (Gao et al., 2017) contributes 3720 official test queries. Every method receives the same frozen ViT-L/14 top-10 within-source pool. The evaluated regime is known-source reranking over benchmark shards or frozen temporal proposals. TVQA queries concatenate the question and five options; moment-localization queries use the official sentence. Retrieval metrics are reported on a 0–100 scale.

Table 2: Split and selection path. "Train-dev" is fixed before training and disjoint by source video or episode. Final-evaluation data and reranker outcomes never select checkpoints, score weights, or audit thresholds.

Corpus	Adaptive training	Selection / threshold fixing	Final evaluation
TVQA	Official train minus train-dev	Train-dev full nDCG; audit thresholds	15,253 test queries
Ego4D-NLQ	Official train minus train-dev	Train-dev full nDCG; audit thresholds	5,127 validation queries
Charades-STA	Official train minus train-dev	Train-dev full nDCG	3720 test queries
SynNearMiss	TVQA-train donors only	No evaluation-driven selection	8,000 perturbations

For TVQA, relevance is binary official-support overlap. For localization, candidate relevance is maximum temporal IoU with any official interval; nDCG uses this graded gain, while MRR and R@1 count a candidate as positive at IoU≥ 0.30. Multiple official intervals reduce by maximum overlap, score ties retain frozen-pool order, and a pool with zero ideal gain receives zero nDCG, MRR, and R@1 rather than entering the audit. Thus full metrics retain pool misses, whereas audit metrics condition on Equation 1.

The comparison roster spans CLIP retrieval (Radford et al., 2021), BM25 (Robertson & Zaragoza, 2009), BERT-base ColBERT late interaction, a BERT-large pairwise cross-encoder, a two-layer 256-dimensional GAT scene-graph matcher, and a four-layer 512-dimensional transformer that jointly encodes the ten candidate descriptor bundles. Adaptive methods train on official retrieval labels and select checkpoints only on the train-derived development split in Table 2. Matched controls additionally consume the same synthetic supervision as TECR. Appendix Table 11 reports the complete roster, and Appendix G gives modalities, candidate representations, parameter counts, supervision, and inference paths. The vanilla cross-encoder is the common reference for gain accounting; Type-Agnostic Conflict is the matched reference for the typed-structure test.

All training and timing use one server with 8×NVIDIA A100 80GB GPUs and CUDA 12.4. The reported reranker-core `ms/query` starts with a materialized dense query representation and precomputed candidate descriptors/graphs; it includes query-to-candidate node alignment and reranking, but excludes dense candidate encoding, query-graph extraction, candidate-state construction, storage, and updates. Timing uses single-query batches, a complete warm-up pass, device synchronization around each query, and the median across the evaluation split. These boundaries are identical in the algorithm, tables, and latency analysis.

5.2 Q1: IS THE GAIN DECOMPOSITION COMPARATOR-ROBUST?

Table 3 reports the development-selected designated checkpoint for each method; Table 5 separately aggregates five independent training seeds under the same selection rule. Table 4 places every primary adaptive method under one reference. All nine rows have positive full and complement gains, while audited conditional gains are 3.20×–5.75× larger than full-set gains. The concentration therefore spans the adaptive roster and all three corpora under one comparison reference.

Table 3: Fixed-pool designated-checkpoint results. Each full cell is nDCG@10/MRR; each audit cell is nDCG@10_{AC}/R@1_{AC} with the exact R@1 count. Parameters exclude frozen extractors. All adaptive rows use identical pools and timing boundaries; OpSup and type-agnostic rows match TECR supervision. Table 5 reports five-seed means and variation.

Method	Params	ms/q	TVQA full | audit	Ego4D full | audit
Cross-Encoder	335M	142	68.3/61.7 | 48.2/39.4 (762/1934)	54.6/47.8 | 35.6/27.4 (188/687)
Cross-Encoder + OpSup	335M	143	68.8/62.0 | 49.8/41.2 (797/1934)	55.2/48.5 | 37.7/29.8 (205/687)
Type-Agnostic Conflict	4.8M	22	68.7/61.9 | 50.5/42.0 (812/1934)	55.3/48.7 | 38.6/30.7 (211/687)
TECR	4.9M	24	69.1/62.8 | 52.4/44.1 (853/1934)	55.8/49.3 | 40.3/32.2 (221/687)

Table 4: Common-reference nDCG@10 gain accounting from five-seed means. Every row uses the vanilla cross-encoder as B; Δ_A and $\Delta_{\bar{A}}$ are audited and complement gains, Γ_A is the pair-specific ACGC ratio, and ρ_A is the audited contribution to the full mean. Values are algebraic decompositions of the rounded reported means.

Corpus	Method	Δ_{all}	Δ_A	$\Delta_{\bar{A}}$	Γ_A	ρ_A
TVQA	Cross-Encoder + OpSup	0.5	1.6	0.34	3.20×	40.6%
	Type-Agnostic Conflict	0.4	2.3	0.12	5.75×	72.9%
	TECR	0.8	4.2	0.31	5.25×	66.6%
Ego4D	Cross-Encoder + OpSup	0.6	2.1	0.37	3.50×	46.9%
	Type-Agnostic Conflict	0.7	3.0	0.34	4.29×	57.4%
	TECR	1.2	4.7	0.66	3.92×	52.5%
Charades	Cross-Encoder + OpSup	0.3	1.5	0.11	5.00×	69.1%
	Type-Agnostic Conflict	0.5	2.3	0.21	4.60×	63.6%
	TECR	1.0	4.1	0.50	4.10×	56.7%

For the primary TECR comparison, the designated checkpoint obtains $\Delta_{\mathrm{all}}^{\mathrm{nDCG}} = 0.8$ on TVQA and 1.2 on Ego4D, while $\Delta_{\mathrm{AC}}^{\mathrm{nDCG}} = 4.2$ and 4.7. TVQA has $p_A = 12.68\%$, a 0.31-point complement gain, and $\rho_{\mathrm{AC}} = 66.6\%$; Ego4D has $p_A = 13.40\%$, a 0.66-point complement gain, and $\rho_{\mathrm{AC}} = 52.5\%$. Official-support recovery at rank 1 rises by 91 and 33 cases, respectively. Absolute paired differences carry the inferential claim; the ratios and shares provide transparent accounting.

5.3 Q2: Does the typed-structure bundle survive matched controls?

The scalar OpSup control measures the value of evidence-change supervision, and the type-agnostic control tests the complete bundle of type identity, relation tags, and constrained alignment at nearly identical capacity. TECR exceeds this stronger control by 1.9 audit nDCG@10 on TVQA (paired 95% CI 1.2–2.7) and 1.7 on Ego4D-NLQ (0.8–2.6). R@1 gains are 2.1 points (1.0–3.3) and 1.5 points (0.4–2.7). Equal descriptors, labels, optimization, and capacity control those alternatives and isolate the combined typed-structure intervention, rather than any one constituent in isolation.

Hierarchical intervals across seeds keep every TECR–vanilla gain positive: TVQA full 0.5–1.1, TVQA audit 3.6–4.8, Ego4D full 0.8–1.6, and Ego4D audit 4.0–5.4. Appendix D reports every seed. The separation between audited and full-set absolute gains therefore persists under training variation.

5.4 Q3: Does gain concentration recur under corpus transfer?

Under the same vanilla cross-encoder reference used for TVQA and Ego4D-NLQ, TECR's Charades full/audited gains are 1.0/4.1 points, ACGC is 4.10×, and the 514 audited queries (13.8%) contribute 56.7% of the rounded full-distribution gain. The complement gains 0.50 points. The separate matched-mechanism comparison gives TECR a 0.5-point overall and 1.8-point audited advantage over Type-Agnostic Conflict (95% CI 0.9–2.8); for this named pair, ACGC is 3.6× and the additive

Table 5: 5-seed nDCG@10 mean ± sample standard deviation. "Full" uses every query; "audit" uses the fixed answer-conflict tag on TVQA and support-conflict tag on Ego4D-NLQ. Each seed is trained independently with identical splits and selection rules.

Method	TVQA full	TVQA audit	Ego4D full	Ego4D audit
Cross-Encoder	68.2±0.25	48.1±0.41	54.5±0.30	35.5±0.48
Cross-Encoder + OpSup	68.7±0.29	49.7±0.47	55.1±0.33	37.6±0.52
Type-Agnostic Conflict	68.6±0.32	50.4±0.43	55.2±0.29	38.5±0.49
TECR	69.0±0.24	52.3±0.39	55.7±0.27	40.2±0.45

Table 6: Charades-STA transfer with the frozen ViT-L/14 top-10 pool and TVQA-only operator pretraining. Metrics aggregate 5 independently trained seeds; the support-conflict audit contains 514 queries.

Method	Full nDCG	MRR	Audit nDCG	Audit R@1
Cross-Encoder	58.0	50.8	38.9	30.0
Cross-Encoder + OpSup	58.3	51.0	40.4	31.3
Type-Agnostic Conflict	58.5	51.3	41.2	32.1
TECR	59.0	52.0	43.0	34.0

share is 49.7%. Charades training uses corpus-specific retrieval labels but no Charades operator labels, extending the support-conflict pattern and matched typed-structure advantage to a third corpus.

5.5 Q4: Which components and conflict families matter?

Table 7 shows a positive TECR ranking advantage within each of the five modeled families on both datasets, with temporal order the smallest effect; the other/compositional intervals cross zero. Removing conflict operators costs 8.6 points, removing operator pretraining costs 4.8, and removing the answer-relevance head costs 3.3. Under equal access to descriptors, labels, and optimization, these results are consistent with a bundle-level typed-comparison explanation. Appendix Table 12 directly tests operator semantics under controlled edits: TECR reaches 71.3% edit recovery versus 42.1% for Scene-Graph Rerank. The controlled edits, component removals, and adjudicated-family gaps connect the learned factors to both intended semantics and natural ranking utility without treating family slices as independent component identifications.

5.6 Q5: What is the reranker-core efficiency tradeoff?

Under the stated materialized-query, precomputed-candidate boundary, TECR's reranker core runs in 24 ms/query, compared with 12 ms/query for embedding-only scoring and 142 ms/query for the BERT-large cross-encoder. The measured core is therefore about 5.9× faster than cross-encoder reranking while improving audit nDCG@10. Type-Agnostic Conflict runs in 22 ms/query, showing the small incremental cost of typed alignment inside this boundary. Appendix H reports the complete timing contract.

6 Discussion

The common-reference matrix shows the breadth of the result: all three adaptive methods improve both audited and complementary populations, yet every audited conditional gain is at least 3.20 times its full-set gain. TECR supplies the largest audited nDCG in all three corpora, while its 66.6%, 52.5%, and 56.7% additive shares quantify the population-level consequence under one consistent reference. Reporting Δ_{all}, Δ_{AC}, $\Delta_{\overline{\mathrm{AC}}}$, ACGC, and ρ_{AC} together separates performance, conditional concentration, and contribution to the mean.

TECR operationalizes this evaluation object through three directed factors. Applicability distinguishes a plausible conflict from arbitrary mismatch, answer relevance focuses the deduction on requested evidence, and reliability suppresses unsupported deductions. Cross-Encoder + OpSup measures the effect of shared evidence-change supervision; Type-Agnostic Conflict then holds descriptors, targets,

Table 7: Real-data conflict localization within the 166 TVQA and 157 Ego4D adjudicated true proposals. Each cell gives count; TECR-minus-supervision-matched Type-Agnostic Conflict audit nDCG@10 gap [five-seed paired-bootstrap 95% CI]. The five modeled families have positive intervals on both datasets.

Conflict family	TVQA	Ego4D-NLQ
Entity substitution	45; 2.4 [1.3, 3.6]	30; 2.2 [0.9, 3.5]
Attribute conflict	25; 1.8 [0.6, 3.0]	20; 1.7 [0.3, 3.2]
Temporal order	24; 0.8 [0.1, 1.6]	28; 0.7 [0.1, 1.3]
Missing region	21; 1.7 [0.4, 3.1]	20; 1.5 [0.1, 2.9]
Missing utterance/audio	21; 2.2 [0.8, 3.7]	21; 2.0 [0.5, 3.5]
Other/compositional	30; 0.5 [-0.4, 1.4]	38; 0.4 [-0.5, 1.3]

Table 8: TVQA audit ablation. Every variant reuses the frozen pool and training labels; the result localizes the contribution of typed conflict scoring.

Variant	nDCG@10$_{\mathrm{AC}}$	Δ	ms/query
TECR	52.4	—	24
− conflict operators	43.8	−8.6	19
− answer-relevance head	49.1	−3.3	23
− operator pretraining	47.6	−4.8	24
− `temp_swap`	50.8	−1.6	23
− `entity_sub`	49.2	−3.2	23

optimization, and capacity fixed while removing the combined typed-structure bundle. Positive paired gaps, component removals, synthetic edit recovery, and natural conflict-family analyses jointly support the bundle-level mechanism.

The established regime is within-source top-10 reranking with fixed extractors and five localized conflict families. Within it, the audit is a gold-support-dependent, high-precision evaluation proposal; localization tags denote support conflict, and rank-1 support recovery is distinct from downstream answer accuracy. Query-bootstrap intervals characterize the observed evaluation queries, and the 24 ms measurement covers reranker-core execution after input materialization. These boundaries preserve the central conclusion: TVQA and Ego4D contain threshold-stable, human-validated conflict populations that carry disproportionate reranking utility, Charades extends the accounting pattern, and typed alignment improves their ordering under matched access and supervision.

7 Conclusion

Video reranking should be evaluated at the evidence state it changes. The Fixed-Pool Answer-Conflict Audit defines that state invariantly across the compared rerankers, and the additive decomposition connects its conditional gain exactly to the full-set mean. Human adjudication, threshold sensitivity, a common-reference matrix, and three corpora establish the concentration pattern; supervision- and capacity-matched controls then show that TECR's combined typed-structure intervention improves the conflict-bearing handoffs beyond an otherwise equivalent type-agnostic model. The result is both an evaluation protocol for locating reranking utility and an efficient mechanism for acting on the evidence conflicts that carry it.

Reproducibility Statement

Sections 3–5 specify the frozen-pool contract, audit rule, model equations, datasets, split exclusions, baseline boundary, selection metric, hardware, and timing contract. Appendix E gives pseudocode, Appendix F gives perturbation and supervision details, Appendix A gives threshold tables, Appendix D gives seed-level results, and Appendix H gives the complete timing procedure.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, organization, terminology, and presentation of author-provided methods, analyses, and results. They did not generate experimental data or independently determine the reported scientific conclusions.

REFERENCES

Zixu Cheng, Yujiang Pu, Shaogang Gong, Parisa Kordjamshidi, and Yu Kong. SHINE: Saliency-aware hierarchical negative ranking for compositional temporal grounding. In *Computer Vision – ECCV 2024*, volume 15077 of *Lecture Notes in Computer Science*, pp. 398–416. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72655-2_23. URL https://doi.org/10.1007/978-3-031-72655-2_23.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A Graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130.

Kevin Flanagan, Dima Damen, and Michael Wray. Moment of untruth: Dealing with negative queries in video moment retrieval. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5336–5345. IEEE, 2025. doi: 10.1109/WACV61041.2025.00521. URL https://doi.org/10.1109/WACV61041.2025.00521.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5267–5275. IEEE, 2017. doi: 10.1109/ICCV.2017.563. URL https://doi.org/10.1109/ICCV.2017.563.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-Retriever: Retrieval-augmented generation for textual graph understanding

and question answering. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/efaf1c9726648c8ba363a5c927440529-Abstract-Conference.html.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxgnnNFvH.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3678. IEEE, 2015. doi: 10.1109/CVPR.2015.7298990. URL https://doi.org/10.1109/CVPR.2015.7298990.

Minjoon Jung, Youwon Jang, Seongho Choi, Joochan Kim, Jin-Hwa Kim, and Byoung-Tak Zhang. Background-aware moment detection for video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8575–8585, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Jung_Background-Aware_Moment_Detection_for_Video_Moment_Retrieval_WACV_2025_paper.html.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR – modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00180. URL https://doi.org/10.1109/ICCV48922.2021.00180.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1167. URL https://aclanthology.org/D18-1167/.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23023–23033. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02205. URL https://openaccess.thecvf.com/content/CVPR2023/html/Moon_Query-Dependent_Video_Representation_for_Moment_Retrieval_and_Highlight_Detection_CVPR_2023_paper.html.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. URL https://arxiv.org/abs/1901.04085.

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. URL https://proceedings.mlr.press/v139/radford21a.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *Computer Vision – ECCV 2018*, volume 11205 of *Lecture Notes in Computer Science*, pp. 690–706. Springer International Publishing, 2018. doi: 10.1007/978-3-030-01246-5_41. URL https://doi.org/10.1007/978-3-030-01246-5_41.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161/.

A AUDIT SENSITIVITY AND ADJUDICATION DETAILS

All sensitivity analyses reuse the identical frozen top-10 lists. Table 9 varies semantic similarity with temporal IoU fixed at 0.30; Table 10 varies temporal IoU with cosine fixed at 0.85. Gaps are TECR minus the BERT-large cross-encoder in nDCG@10_{AC}, intervals use paired query bootstrap with 10000 resamples, and ACGC follows Equation 3 using full-distribution gaps of 0.8 for TVQA and 1.2 for Ego4D-NLQ.

Adjudication protocol. For each dataset, the audit-validity sample contains 200 randomly selected tagged queries and 100 eligible non-tagged queries nearest to the fixed two-dimensional boundary. Boundary distance is the minimum normalized change in cosine or temporal IoU required to satisfy Equation 1, with the other coordinate held fixed; frozen-pool rank resolves ties. When several

Table 9: Cosine-threshold sensitivity at temporal IoU 0.30. The center row is the fixed-before-evaluation operating point; neighboring rows quantify stability around it.

Dataset	τ_s	Queries	Gap	95% CI	ACGC
TVQA	0.80	2781	3.5	2.9–4.1	4.38×
TVQA	0.825	2316	3.9	3.3–4.5	4.88×
TVQA	0.85	1934	4.2	3.6–4.8	5.25×
TVQA	0.875	1519	4.4	3.7–5.1	5.50×
TVQA	0.90	1107	4.3	3.4–5.2	5.38×
Ego4D	0.80	1009	4.0	3.2–4.8	3.33×
Ego4D	0.825	836	4.3	3.5–5.1	3.58×
Ego4D	0.85	687	4.7	3.9–5.5	3.92×
Ego4D	0.875	531	4.8	3.8–5.8	4.00×
Ego4D	0.90	397	4.6	3.4–5.8	3.83×

Table 10: Temporal-IoU sensitivity at cosine 0.85. ACGC remains large while looser overlap thresholds expand the audited population.

Dataset	τ_o	Queries	Gap	95% CI	ACGC
TVQA	0.20	1497	4.6	3.9–5.3	5.75×
TVQA	0.25	1719	4.4	3.8–5.0	5.50×
TVQA	0.30	1934	4.2	3.6–4.8	5.25×
TVQA	0.35	2163	4.3	3.7–4.9	5.38×
TVQA	0.40	2391	3.9	3.4–4.4	4.88×
Ego4D	0.20	506	5.0	4.0–6.0	4.17×
Ego4D	0.25	596	4.8	3.9–5.7	4.00×
Ego4D	0.30	687	4.7	3.9–5.5	3.92×
Ego4D	0.35	781	4.5	3.8–5.2	3.75×
Ego4D	0.40	879	4.6	3.9–5.3	3.83×

candidates qualify, annotators inspect the canonical $d^\star$ defined in Section 3, so the reported precision is at the query–candidate level rather than an average over every distractor.

Two annotators independently inspect the query, official option or interval, official-support shard, and $d^\star$ while blinded to model identity and reranked scores. For TVQA they label whether the candidate supports a non-gold option; for localization they label whether it is semantically plausible for the query but conflicts with official temporal support. They then assign one conflict family or other/compositional and may record insufficient evidence. Disagreements are resolved by a third adjudicator. Exact counts, Wilson intervals, and Cohen's κ appear in Table 1. A schematic true proposal pairs a visually plausible candidate with the wrong queried entity or interval; a false proposal is close in the frozen embedding yet lacks enough candidate evidence to establish that conflict. These labels evaluate the automatic proposal rule rather than redefining it after observing model outcomes.

B Real-data conflict taxonomy

Table 7 places the complete adjudicated taxonomy in the main paper. The five modeled families cover 136 of 166 adjudicated TVQA answer-conflict cases and 119 of 157 Ego4D support-conflict cases. Positive paired gaps within every modeled family are consistent with family-localized ranking utility. The smaller temporal effect and intervals crossing zero for other/compositional cases delimit the evidence for the current operator vocabulary.

C Full benchmark and mechanism results

The two operator values beyond Table 12 are 70.4% region-removal and 78.3% audio-drop accuracy for TECR, versus 39.8% and 45.2% for Scene-Graph Rerank. Aggregate recovery is computed over

Table 11: Full measured baseline roster under the same fixed top-10 pools. Metrics are on a 0–100 scale. The reconciled Ego4D R@1 entries report exact binary-query counts.

Method	ms/q	nDCG@10	MRR	nDCG@10$_{\mathrm{AC}}$	R@1$_{\mathrm{AC}}$
TVQA					
CLIP-Retrieve	12	62.4	55.1	41.3	32.8
BM25	8	54.8	47.3	38.6	29.5
ColBERT	18	65.1	58.3	44.7	36.2
Cross-Encoder	142	68.3	61.7	48.2	39.4
Scene-Graph Rerank	35	64.7	57.9	45.1	36.8
Memory-Transformer	87	66.9	60.2	46.8	38.1
TECR	24	69.1	62.8	52.4	44.1
TECR without conflicts	19	65.4	58.6	43.8	35.6
Ego4D-NLQ					
CLIP-Retrieve	12	48.7	41.2	29.1	21.4
BM25	8	39.2	33.7	25.4	18.7
ColBERT	18	51.3	44.1	32.8	24.1
Cross-Encoder	142	54.6	47.8	35.6	27.4
Scene-Graph Rerank	35	50.8	43.5	33.2	25.0
Memory-Transformer	87	53.1	46.2	34.9	26.8
TECR	24	55.8	49.3	40.3	32.2
TECR without conflicts	19	49.9	42.8	31.5	23.7

Table 12: SynNearMiss mechanism results with exact operator labels. This in-family diagnostic tests whether learned heads recover their intended conflict semantics.

Method	Edit recovery	nDCG@10	R@1	Entity	Temporal	Attribute
CLIP-Retrieve	—	51.4	38.7	—	—	—
Cross-Encoder	—	62.8	49.3	—	—	—
Scene-Graph Rerank	42.1	58.6	44.8	48.3	35.7	41.9
TECR	71.3	72.1	61.4	76.8	64.2	72.1

all perturbations rather than as an unweighted operator macro-average; the resulting 71.3% therefore reflects the benchmark's class frequencies while retaining a difficult temporal class.

Across the five similarity bins (0.5–0.6 through 0.9–1.0), gains are Cross-Encoder [3.1, 3.5, 4.2, 5.1, 5.8], Type-Agnostic Conflict [3.0, 3.8, 5.1, 7.0, 8.1], and TECR [3.2, 4.1, 5.7, 8.3, 10.4].

D SEED-LEVEL UNCERTAINTY

Table 13: TVQA nDCG@10 for five independent training seeds. Each bracket lists seeds in the fixed prespecified order.

Method	Full-distribution seeds	Audit seeds
Cross-Encoder	[68.2, 67.9, 68.5, 68.0, 68.4]	[48.1, 47.6, 48.6, 47.8, 48.4]
Cross-Encoder + OpSup	[68.70, 68.34, 69.06, 68.51, 68.89]	[49.70, 49.10, 50.30, 49.43, 49.97]
Type-Agnostic Conflict	[68.6, 68.2, 69.0, 68.4, 68.8]	[50.40, 49.85, 50.95, 50.15, 50.65]
TECR	[69.00, 68.69, 69.31, 68.86, 69.14]	[52.30, 51.80, 52.80, 52.07, 52.53]

Seeds are fixed before training. Every adaptive method uses the same source-disjoint training/development partition and early-stopping rule; checkpoint selection maximizes full-distribution development nDCG@10. Per seed, paired query bootstrap estimates the within-seed difference. A hierarchical bootstrap first samples seeds and then queries within the sampled seed to form the aggregate intervals reported in Section 5.3.

E ALGORITHM AND TRAINING DETAILS

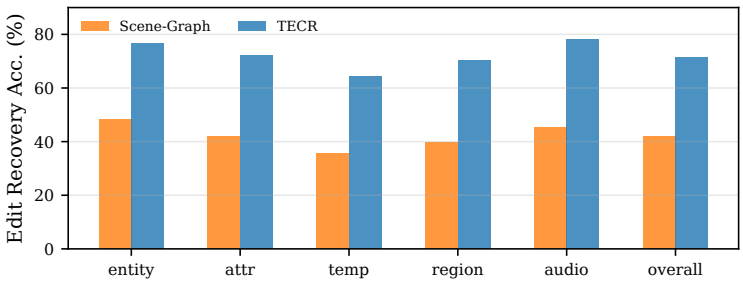


Figure 3: Per-operator edit-recovery accuracy on SynNearMiss. Temporal swaps remain the hardest class; entity substitutions and audio omissions are the strongest, and the remaining families fall between them.

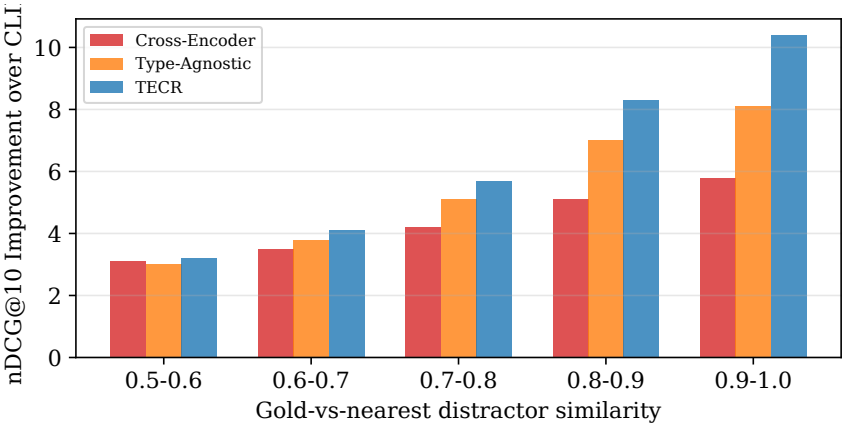


Figure 4: TVQA nDCG@10 gain over CLIP retrieval by gold–distractor similarity. Separation grows as dense similarity collapses answer-incompatible alternatives, consistent with the fixed audit's ACGC pattern.

F Training scope and perturbation recipes

Operator pretraining begins from a gold TVQA training shard and samples one eligible node or edge. `entity_sub` replaces an entity and its mentions with a same-type donor from the training split; `attr_flip` changes a discrete attribute on an aligned entity or region; `temp_swap` exchanges adjacent action or utterance order; `region_rem` removes one grounded region and incident edges; and `audio_drop` removes one utterance and speaker link. The applicability target identifies the sampled valid operator; the answer-relevance target is one exactly when the modified element overlaps query-anchored gold evidence. The margin-ranking target orders the gold shard above its evidence-changing perturbation, and the reliability output is gated by whether the corresponding operator deduction reduces that ranking residual. Invalid alignment, identity edits, and candidates outside the source episode are rejected. Donor pools exclude TVQA validation/test episodes, all Ego4D and Charades videos, and every real evaluation shard.

All three supervised rerankers use the same retrieval examples, perturbation pairs, operator classes, evidence-change targets, optimizer schedule, update count, source-disjoint development split, and full-distribution selection metric. Cross-Encoder + OpSup maps operator labels to auxiliary scalar-token heads; Type-Agnostic Conflict retains the same head and loss structure as TECR while jointly erasing type identities, relation identities, and constrained alignment. The shared multitask objective is $\mathcal{L}_{\text{op}} + 0.5\mathcal{L}_{\text{answer}} + \mathcal{L}_{\text{rank}}$.

Table 14: Ego4D-NLQ nDCG@10 for the same five-seed protocol.

Method	Full-distribution seeds	Audit seeds
Cross-Encoder	[54.50, 54.11, 54.89, 54.33, 54.67]	[35.50, 34.88, 36.12, 35.22, 35.78]
Cross-Encoder + OpSup	[55.10, 54.67, 55.53, 54.91, 55.29]	[37.60, 36.93, 38.27, 37.30, 37.90]
Type-Agnostic Conflict	[55.20, 54.82, 55.58, 55.04, 55.36]	[38.50, 37.87, 39.13, 38.21, 38.79]
TECR	[55.70, 55.35, 56.05, 55.54, 55.86]	[40.20, 39.62, 40.78, 39.94, 40.46]

Algorithm 1 Typed Evidence Conflict Reranking

Require: Query q, frozen candidates C_q, typed candidate states, operator set $\mathcal{O}$

Read materialized dense query representation and precomputed candidate states
Construct query graph outside the reported reranker-core timing region
for each candidate $m_i \in C_q$ **do**
  Align nodes by type; map unmatched nodes to NULL
  Compute shared aligned representation h_{qi} once
  Compute support $U(q, m_i)$
  for each $o \in \mathcal{O}$ **do**
    Compute conflict triple (a_o, r_o, c_o) from h_{qi}
  end for
  $B(q, m_i) \leftarrow |\mathcal{O}|^{-1} \sum_o a_o r_o c_o$
  $s(q, m_i) \leftarrow (1 - \alpha - \beta)\widetilde{s}_{\text{emb}} + \alpha U + \beta(1 - B)$
end for
Return C_q sorted by decreasing $s(q, m_i)$

G Architecture and supervision parity

The parity boundary separates frozen candidate generation from adaptive scoring. Query graph extraction and candidate state construction precede timing for both conflict models; serialized descriptor preparation precedes timing for cross-encoders; and all reranker-core numbers include the model-specific processing, candidate comparison, and score sorting that follow those materialized inputs.

H Latency protocol and operating points

Timing runs on one 8×NVIDIA A100 80GB server under CUDA 12.4, with one query per batch. A complete untimed pass warms model weights, kernels, and cached candidate state. Each measured query uses device synchronization immediately before and after the timed region; reported values are medians across the official evaluation split. The region begins with a materialized dense query representation and candidate state, includes model-specific processing required after those inputs, node alignment where applicable, scoring of all ten candidates, and sorting, and ends after ranked indices are materialized. Candidate encoding, query-graph extraction, graph construction, graph storage, and memory-bank updates remain outside this reranker-core comparison.

Table 15: Reproducible adaptive-reranker definitions. Parameter counts exclude frozen extractors and include every trained scoring component. All methods receive the same top-10 pool.

Method	Candidate representation	Trainable params	Supervision	Inference path
ColBERT	Shard transcript and labels	110M	Retrieval labels	BERT-base token encoding and late interaction
Cross-Encoder	Frozen descriptor bundle serialized with query	335M	Retrieval labels	Ten independent BERT-large pair scores
Cross-Encoder + OpSup	Same serialized bundle	335M	Retrieval + shared operator targets	Same pair scores plus auxiliary training heads
Scene-Graph Rerank	Typed local graph	3.2M	Retrieval labels	Two-layer 256-d four-head GAT and pooled match
Memory-Transformer	Ten descriptor bundles jointly	26.7M	Retrieval labels	Four-layer 512-d eight-head candidate attention
Type-Agnostic Conflict	Same descriptor bundle as TECR	4.8M	Retrieval + shared operator targets	Unconstrained alignment, shared attention, five heads
TECR	Typed local graph	4.9M	Retrieval + shared operator targets	Type-constrained alignment, shared attention, five heads