Does AI Assistance Preserve or Collapse Disagreement? A Study of Pre-Annotations in Ambiguous Video Labeling

Juan Gutiérrez [1 2] **Victor Gutiérrez-García** [2] **José Luis Blanco-Murillo** [1 2]

Abstract

AI-generated Pre-Annotations can accelerate video labeling, but they may also anchor annotators to model priors and suppress disagreement that is valuable for pluralistic dataset construction. We study this tradeoff in ambiguous temporal video annotation, where annotators choose event boundaries and assign context-dependent labels such as "normal" or "abnormal." We introduce a controlled audit protocol that separates annotation cost, consensus alignment, inter-annotator consistency, temporal-boundary variation, semantic-label variation, latent-space standardization, and edit behavior.

In a counterbalanced pilot study with 18 annotators and 180 annotation sessions, a fixed CLIP-based Pre-Annotation engine reduced mean annotation time by 23.11%; 72% of annotators were faster with assistance, with a median per-annotator gain of 35%. Assistance increased inter-annotator consistency and CLIP-space standardization, while semantic-label entropy changed only slightly in aggregate. A six-annotator human consensus diagnostic showed comparable alignment across conditions (AMI ≈ 0.64 in both conditions), but this diagnostic is self-including and should be interpreted as descriptive rather than as independent evidence that model influence is absent. Overall, the results suggest that Pre-Annotations acted mainly as editable temporal scaffolds in this pilot setting, while leaving open the possibility of subtler semantic anchoring effects. We contribute an audit framework and anonymized interaction traces for studying when AI-assisted annotation preserves, reshapes, or collapses human disagreement.

[1]Information Processing and Telecommunications Center [2]ETSI de Telecomunicación, Universidad Politécnica de Madrid, Spain. Correspondence to: <juan.gutierrez@upm.es>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Pluralistic alignment shifts the goal of AI development from optimizing toward a single aggregated preference to representing, preserving, and reasoning about the diversity of human judgments. This shift is especially important in data collection pipelines, where disagreement is often collapsed through majority voting, adjudication, or averaging (Uma et al., 2021; Davani et al., 2022). In subjective or context-dependent tasks, disagreement can reflect different interpretations, experiences, assumptions, or values rather than simple annotation error (Fazelpour & Fleisher, 2025). Recent work argues that alignment datasets and benchmarks should preserve richer distributions of human judgments rather than assuming a unique ground truth (Gordon et al., 2022; Sorensen et al., 2024).

We study this issue in dense temporal video annotation. Unlike static image labeling, video annotation requires annotators to decide when events begin and end, how temporally extended activities should be grouped, and how observed behavior should be classified. Decisions are often ambiguous: event boundaries can be gradual, multiple actors may interact, and the meaning of an event depends on its context. In anomaly-oriented video labeling, the distinction between "normal" and "abnormal" behavior is especially sensitive to interpretation, social context, and the operational definition given to annotators (Sultani et al., 2018; Wu et al., 2024).

AI-assisted annotation alters this decision process. Model-generated Pre-Annotations can reduce the mechanical burden of drawing segments from scratch, shifting the annotator's role from creation to verification and refinement. Yet this same assistance can also act as an anchor. If annotators accept proposed boundaries or labels too readily, the resulting dataset may appear cleaner while losing disagreement that would otherwise be informative for pluralistic modeling. The key methodological challenge is therefore not simply to measure whether assistance improves speed or agreement, but to determine what kind of agreement it increases.

We therefore ask how AI-generated Pre-Annotations change the distribution of human annotations in ambiguous temporal video labeling. Do they merely reduce mechanical effort and temporal jitter, or do they anchor annotators toward

model priors and suppress meaningful disagreement? We answer this question with a controlled A/B study in which 18 annotators label high-ambiguity video clips with and without AI-generated Pre-Annotations. The Pre-Annotation engine is used as a fixed intervention that supplies temporally localized suggestions. We instrument the annotation interface to collect final labels, timing information, model suggestions, and interaction traces.

Our evaluation addresses three axes that are often conflated. First, we measure annotation cost through time-on-task. Second, we measure consensus alignment by comparing each annotation to a human consensus diagnostic, while avoiding the assumption that this reference is objective ground truth. Third, we measure disagreement and standardization through pairwise inter-annotator consistency and latent-space structure. This separation allows us to distinguish a reduction in arbitrary temporal boundary jitter from a collapse of meaningful annotator variation.

Our contributions are:

- **A disagreement-centered framing of AI-assisted annotation**: we treat Pre-Annotations as behavioral interventions on the distribution of human annotations, rather than only as tools for improving throughput.
- **A controlled audit protocol** that separates annotation cost, consensus alignment, inter-annotator consistency, temporal-boundary variation, semantic-label variation, latent-space standardization, and edit-trace behavior.
- **A counterbalanced pilot study with 18 annotators and 180 annotation sessions** on ambiguous temporal video labeling.
- **Empirical evidence that assistance improves speed and increases standardization**: Pre-Annotations reduce mean annotation time by 23.11% and increase pairwise consistency. A self-including human consensus diagnostic shows comparable aggregate alignment across conditions, which we report as descriptive evidence rather than as an independent test of whether model influence is absent.
- **A planned release of anonymized interaction traces and analysis code** to support follow-up work on pluralistic data collection, anchoring effects, and human–AI annotation workflows.

2. Related Work

Pluralistic alignment and annotation disagreement. Pluralistic alignment studies how AI systems can account for heterogeneous human values, preferences, and judgments rather than optimizing toward a single averaged target (Sorensen et al., 2024). Annotation disagreement is central to this problem. In many pipelines, multiple labels are reduced to a single value through majority vote. However, prior work shows that disagreement can be systematic and informative in subjective tasks (Uma et al., 2021; Davani et al., 2022). Methods such as jury learning preserve dissenting voices and enable downstream models to reason over distributions of human judgment (Gordon et al., 2022). Our work studies an upstream stage of this pipeline: how does AI assistance during annotation reshape the disagreement that pluralistic modeling aims to preserve?

AI-assisted annotation and human verification. AI-assisted annotation is commonly motivated by efficiency. A model proposes labels, regions, or temporal segments, and humans correct them through a verification interface. Prior work has explored verification-centric labeling, interaction-efficient interfaces, and video-specific annotation tools (Papadopoulos et al., 2016; 2017; Kuznetsova et al., 2021; Deng et al., 2021; Feng et al., 2023; Shrestha et al., 2023). Production tools such as CVAT and Label Studio have made model-assisted labeling a practical workflow for dataset construction (Tkachenko et al., 2020; CVAT.ai Corporation, 2023). These systems are usually evaluated through throughput or final-label quality. We evaluate an additional question: how suggestions alter (improve or degrade) the structure of annotator disagreement.

Ambiguous video labeling. Dense temporal video annotation is a useful setting for studying disagreement because annotators must make both temporal and semantic decisions. Open-world anomaly datasets such as UCF-Crime contain untrimmed surveillance-style footage with low resolution, unconstrained viewpoints, gradual event transitions, and ambiguous behavioral categories (Sultani et al., 2018). Recent video anomaly detection work has adapted vision–language models to these datasets (Wu et al., 2024), but the human annotation process remains difficult to evaluate since no single frame-perfect ground truth captures all reasonable interpretations. We use this ambiguity as the target condition for our audit.

Vision–language models as annotation scaffolds. Vision–language models such as CLIP provide transferable visual and textual embeddings that can be adapted to video tasks with modest computation (Radford et al., 2021). Efficient CLIP-style video learners use frame-level encoding, temporal pooling, and targeted fine-tuning to produce video-text representations (Rasheed et al., 2023; Wang et al., 2023). Our Pre-Annotation engine follows this resource-conscious approach and combines a CLIP-style encoder with hierarchical spherical k-means (Hornik et al., 2012). The engine supplies a stable source of suggestions for the human-subjects audit.

3. Evaluation and Analysis

Figure 1 summarizes the audit pipeline. We collect annotations in Assisted and Unassisted conditions, preserve disaggregated interaction traces, and evaluate the resulting data along three axes: annotation cost, consensus alignment, and disagreement/standardization.

3.1. Materials and Model Sanity Check

The video clips used in this study were sampled from a held-out pool of open-world videos (UCF-Crime (Sultani et al., 2018)) containing both normal and abnormal events. The Pre-Annotation engine was trained and validated on the video-language datasets described in Appendix A.

The selected fixed encoder achieved 89.68% Acc@1 and 99.64% Acc@5 in video-text matching. This result is used as a sanity check to ensure that the Pre-Annotation source produces semantically plausible video-text representations. It does not, by itself, validate the temporal quality of the proposed segments.

Pre-Annotation intervention. The Assisted condition displayed model-generated temporal segments with proposed semantic labels and binary normal/abnormal flags. Annotators could accept, edit, relabel, split, merge, or remove these suggestions before submission. The same fixed Pre-Annotation engine was used for all Assisted sessions. The engine was not updated during the study and did not adapt to annotator behavior. This makes the model output a stable experimental intervention rather than an adaptive collaborator. Screenshots and implementation details of the annotation interface are provided in Appendix B.

3.2. Disaggregated Annotations and Interaction Traces

We collected anonymized annotator-level traces for all 180 annotation sessions. The traces include final temporal segments, class labels, binary normal/abnormal flags, model Pre-Annotations, session-level timing, and edit histories. We preserve these disaggregated records because pluralistic analysis requires more than an aggregated final label. It requires visibility into where annotators converged, when they disagreed, and how model suggestions changed the annotation process.

Study design. The study used a within-subject, counterbalanced design. Each of the 18 annotators labeled ten videos: five in the Assisted condition, where model-generated Pre-Annotations were visible, and five in the Unassisted condition, where annotators labeled raw footage from scratch. Annotators did not label the same video twice. Across the study, this produced 180 annotation sessions over 30 videos, with six independent annotations per video. Condition order and video assignment were counterbalanced to reduce order and learning effects. Each annotation session is therefore nested within both an annotator and a video, and we treat annotator identity and video identity as repeated-measures factors in the analysis. Additional documentation of the annotation task, annotator recruitment, platform choices, and participant-protection considerations is provided in Appendix C.

Table 1. Summary of the disaggregated annotation traces collected during the controlled study.

Metric	Value
Annotation and Segments	
Total Annotators	18
Total Annotation Sessions	180
Total Labeled Segments	699
Semantics	
Mean Classes per Video	3.01 ± 0.95
Segments with Binary Event Flag	670
Abnormal Event Segments	323
Normal Event Segments	347

Table 1 summarizes the collected traces. The released artifact is intended to support follow-up analyses of efficiency, revision behavior, anchoring, and disagreement preservation in AI-assisted annotation workflows.

3.3. Metrics

Our evaluation separates the annotation cost, aggregate consensus alignment, inter-annotator consistency, and the structure of disagreement. This separation is important because higher agreement can reflect either a desirable reduction of arbitrary boundary jitter or an undesirable collapse of meaningful annotator variation.

Annotation cost. We measure time-on-task in seconds from initial video load to final submission. We report session-level means and standard deviations, per-annotator percentage changes between conditions, and a mixed-effects estimate of the Assisted-condition effect on log time. Since annotation sessions are repeated within annotators and videos, we treat the timing results as repeated-measures data rather than as independent observations.

Consensus alignment. For each video, we construct a human Consensus Reference from the six independent annotations assigned to that video. We compare individual annotations to this reference using Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), and V-measure (Rosenberg & Hirschberg, 2007; Vinh et al., 2010). These metrics compare induced temporal label partitions

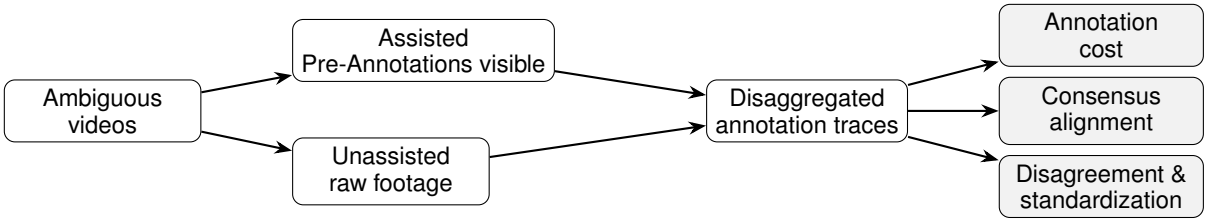


Figure 1. Pluralistic annotation audit framework. We evaluate whether Pre-Annotations reduce effort while preserving, reshaping, or collapsing human disagreement.

while accounting for chance structure in cluster assignments. The Consensus Reference is used only as an aggregate descriptive diagnostic, not as objective ground truth. Because the evaluated annotation is included in the six-annotator reference, this diagnostic is self-including and may inflate alignment values. We therefore do not interpret comparable consensus alignment as proof that anchoring is absent; instead, we interpret it jointly with pairwise consistency, boundary dispersion, semantic-label entropy, latent-space standardization, and edit traces.

Temporal partition construction. To compute AMI, NMI, V-measure, and entropy-based diagnostics, we convert each temporal annotation into a discrete label sequence over fixed temporal bins of width $\Delta = 25$ seconds. Each bin receives the semantic class label, binary normal/abnormal flag, or background label corresponding to the segment with the largest temporal overlap with that bin. If no segment overlaps the bin, the bin is assigned the background label. If an annotator creates overlapping segments with different labels, the bin is assigned the label of the segment with the largest overlap; remaining ties are resolved by selecting the segment with the earliest start time. These bin-level sequences induce temporal label partitions that can be compared across annotators and against the Consensus Reference. Unless otherwise stated, agreement metrics are computed per video and summarized across videos. We did not perform a bin-width sensitivity analysis in this pilot, so agreement and entropy values should be interpreted relative to this fixed discretization choice.

Avoiding consensus circularity. Consensus-alignment metrics are computed against a six-annotator human Consensus Reference for each video. For every video v, the reference is constructed from all six independent annotations assigned to that video, aggregating their temporal label partitions into a single human consensus diagnostic. This construction intentionally describes alignment with the observed annotation distribution, but it also introduces circularity: when an individual annotation is evaluated, that annotation contributes to the reference against which it is compared. As a result, consensus-alignment values are likely optimistic and may obscure subtle differences between conditions. We therefore treat consensus alignment as a descriptive diagnostic only. The central interpretation of the study rests on the pattern across diagnostics: timing, pairwise consistency, boundary dispersion, semantic-label entropy, latent-space standardization, and edit behavior.

Inter-annotator consistency. We measure pairwise agreement among annotators within each condition. Higher pairwise consistency can indicate useful coordination around event boundaries, but it can also indicate disagreement collapse. We therefore interpret pairwise consistency jointly with consensus alignment, boundary-level variation, semantic-label variation, and interaction traces.

Boundary-level disagreement. To distinguish boundary standardization from semantic convergence, we separately measure temporal boundary variation. Because annotators may produce different numbers of segments, we first match segments within each video using maximum temporal overlap subject to label compatibility. For each matched event, we compute the dispersion of start and end times across annotators:

$$\sigma_{\text{start}}(v, e) = \text{StdDev}\{s_{a,v,e}\}_{a=1}^{A}, \tag{1}$$

$$\sigma_{\text{end}}(v, e) = \text{StdDev}\{t_{a,v,e}\}_{a=1}^{A}, \tag{2}$$

where $s_{a,v,e}$ and $t_{a,v,e}$ are the start and end times assigned by annotator a to event e in video v. Lower boundary dispersion in the Assisted condition indicates temporal standardization, but not necessarily semantic agreement. Unmatched segments are reflected separately through segment-count variation.

Semantic-label disagreement. We measure variation in event labels and binary normal/abnormal flags separately from boundary placement. For each video and condition, we compute the entropy of the empirical label distribution:

$$H_v = -\sum_{c \in \mathcal{C}} p_v(c) \log p_v(c). \tag{3}$$

where $p_v(c)$ is the proportion of temporal bins in video v assigned to label c across annotators in that condition. This diagnostic helps distinguish reduced boundary jitter from reduced diversity of semantic interpretation.

Latent-space standardization. We compute the Silhouette Score of human-labeled segments using CLIP image embeddings and cosine distance (Shahapure & Nicholas, 2020; Radford et al., 2021). Because the same encoder family contributes to the Pre-Annotation engine, this metric measures coherence under the model's representational geometry. We therefore treat it as a standardization diagnostic rather than as an independent measure of semantic correctness.

Anchoring diagnostics. For Assisted sessions, we additionally compute interaction-level diagnostics from edit traces: the fraction of Pre-Annotation segments accepted unchanged, boundary edited, split, merged, deleted, or relabeled; the mean boundary displacement from the model suggestion to the final human segment; and the label-flip rate between the model suggestion and the final annotation. These metrics directly test whether annotators used Pre-Annotations as editable scaffolds or accepted them with minimal revision.

Statistical analysis. The session results are not independent. We analyze the effects of conditions using repeated-measures estimates. For timing, we fit a mixed-effects model of the form

$$\begin{aligned}\log(\text{time}_{a,v}) = \beta_0 &+ \beta_1 \,\text{Assisted}_{a,v} \\ &+ \beta_2 \,\text{Order}_{a,v} + \beta_3 \,\text{Duration}_v \\ &+ u_a + w_v + \epsilon_{a,v}. \end{aligned} \tag{4}$$

where u_a is a random intercept for annotator and w_v is a random intercept for video. We report the Assisted-condition coefficient as the primary repeated-measures estimate of annotation-time reduction. For agreement, disagreement, and standardization metrics, we report video-level summaries. Session-level t-tests and Wilcoxon tests are included only as descriptive checks, not as the primary inferential evidence, because they do not fully account for repeated measurements within annotators and videos.

3.4. Annotation Cost

The first audit axis is annotation cost. We tracked the duration of the task, from initial video load to final submission, across all 180 sessions. Figure 2 shows the time-on-task distributions for the Assisted (with-) and Unassisted (without-Pre-Annotations) conditions. Table 2 reports the results on both parametric and non-parametric session-level statistical tests, which indicate a reliable timing difference in the collected sessions.

As shown in Table 2, the Assisted condition reduced the mean annotation time from 190.44 seconds to 146.42 seconds, corresponding to a 23.11% reduction in mean time-on-task. Normalizing within annotator, 72% of participants were faster with Pre-Annotations, with a median per-annotator gain of 35%. A mixed-effects model with annotator and video random intercepts produced an Assisted-condition coefficient of approximately $\hat{\beta}_1 = -0.263$ on log time, corresponding to an estimated 23.1% reduction in annotation time. The session-level t-test and Wilcoxon test are reported only as descriptive checks because they do not fully account for repeated measurements within annotators and videos.

The per-annotator comparison in Figure 3 shows that the efficiency benefit is not driven solely by a small number of sessions. On these axes, the x-coordinate is Assisted time and the y-coordinate is Unassisted time; therefore, points above the diagonal indicate faster completion with Pre-Annotations. For each annotator, we removed the first

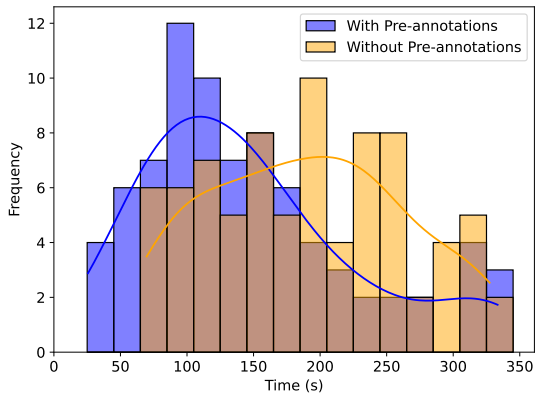
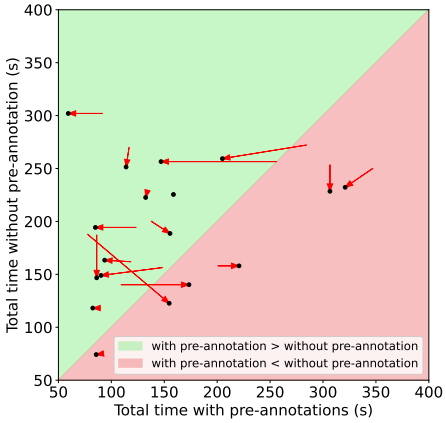


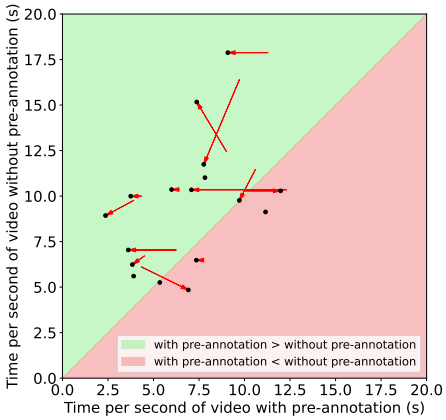
Figure 2. Distribution of annotation times with and without Pre-Annotations. Bin width is 20 seconds.

Table 2. Annotation-time results. Time is measured in seconds from initial video load to final submission. Uncertainty is reported as approximate 95% confidence intervals. The mixed-effects estimate is reported on log time and transformed into an implied percentage reduction.

Condition	**Mean**	**Std.**	**95% CI**
Assisted	146.42	79.28	± 16.60
Unassisted	190.44	72.63	± 15.22
Effect estimates			
Mean reduction	23.11%		
Mean difference	-44.02 ± 22.44 s		
Annotators faster with assistance	72%		
Median per-annotator gain	35%		
Repeated-measures estimate			
Mixed-effects coefficient, $\hat{\beta}_1$	-0.263 [-0.397, -0.129]		
Implied time reduction	23.1% [12.1%, 32.8%]		
Wald test	$z = -3.86$, $p < 0.001$		
Session-level descriptive tests			
t-test	$t = -3.64$, $p = 0.0004$		
Wilcoxon	$W = 853.0$, $p = 0.0002$		



(a) Normalized annotation time



(b) Normalized shift after training videos

Figure 3. Per-annotator efficiency effects before and after removing early familiarization trials. Each red arrow represents the change for an annotator after excluding the first four training videos. Times are normalized by video duration. Points in the assisted-faster region indicate annotators who completed annotations faster with Pre-Annotations.

four videos they annotated in chronological order and recomputed the per-annotator normalized timing comparison. The normalized comparison after the initial familiarization videos shows a similar pattern, suggesting that annotators became more efficient at reviewing and editing suggestions once they were familiar with the interface.

Annotator feedback suggests that Pre-Annotations mainly reduced the time spent deciding on an initial segmentation structure. In some cases, annotators reported that the suggested segments were sufficiently close to require only review and minor correction. In other cases, the suggestions required substantial editing. This variation matters for pluralistic annotation: the aim is not to make humans accept outputs quickly, but to reduce mechanical effort while preserving the opportunity for correction and disagreement.

Table 3. Consensus alignment and disagreement diagnostics. Positive deltas indicate larger assisted-condition values.

Diagnostic	Assisted	Unassisted	Δ
Consensus AMI	0.64	0.63	+0.01
Consensus NMI	0.64	0.64	0.00
Consensus V-measure	0.64	0.64	0.00
Pairwise AMI	0.67	0.62	+0.05
Pairwise NMI	0.67	0.62	+0.05
Pairwise V-measure	0.67	0.62	+0.05

3.5. Consensus Alignment and Inter-Annotator Consistency

The second audit axis asks whether AI assistance changes semantic judgment. In pluralistic annotation, higher agreement is ambiguous: it can indicate useful reduction of boundary jitter, but it can also indicate homogenization around model priors. We therefore separate two measurements: alignment with a six-vote human Consensus Reference and pairwise consistency among annotators within each condition.

Table 3 shows that Assisted and Unassisted annotations have nearly identical alignment with the Consensus Reference. The Assisted condition obtains AMI 0.64, while the Unassisted condition obtains AMI 0.63. NMI and V-measure are also comparable across conditions. However, this consensus diagnostic is self-including: each evaluated annotation contributes to the six-annotator reference against which it is compared. These values should therefore be interpreted as descriptive alignment with the observed annotation distribution, not as an independent validation that assistance preserves human judgment or eliminates anchoring.

Pairwise inter-annotator consistency is higher in the Assisted condition. Assisted annotators reached AMI 0.67, compared to AMI 0.62 for the Unassisted condition. This indicates standardization: annotators who see the same suggestions converge more strongly with each other. However, higher consistency is not, by itself, evidence that meaningful disagreement has been preserved. It may reflect a useful reduction of arbitrary temporal jitter, but it may also reflect anchoring toward model-proposed boundaries or labels. We therefore treat the increase in pairwise agreement as a signal requiring decomposition into boundary-level and semantic-label components, rather than as a standalone success criterion.

3.6. Boundary and Semantic Disagreement Decomposition

To better interpret the increase in pairwise consistency, we decompose disagreement into temporal boundary variation and semantic-label variation. Boundary variation captures

Table 4. Disagreement decomposition. Boundary-level metrics distinguish temporal standardization from semantic-label convergence. Values are computed per video and summarized across videos.

Diagnostic	Assisted	Unassisted	Difference
Start-boundary dispersion	0.80	1.10	-0.30
End-boundary dispersion	1.21	2.10	-0.89
Segment-count variance	1.26	0.72	+0.54
Class-label entropy	1.17	1.18	-0.01
Normal/abnormal entropy	0.60	0.64	-0.04

differences in the placement of annotators' event start and end times. Semantic-label variation captures differences in the class labels and normal/abnormal flags assigned to the segments.

This decomposition is important because the pluralistic impact of a disagreement depends on its source. Reduced boundary jitter may be desirable when annotators identify the same event but differ by a few frames or seconds. Reduced semantic-label entropy, by contrast, may indicate that model suggestions suppress alternative interpretations of ambiguous events.

Table 4 reports three additional diagnostics: boundary dispersion, segment-count variation, and label entropy. Boundary dispersion measures the standard deviation of matched start and end times across annotators. Segment-count variation measures whether assistance causes annotators to produce a more uniform number of segments. Label entropy measures whether assistance reduces diversity in semantic labels or binary event flags.

Interestingly, assistance reduced boundary dispersion for matched events but increased segment-count variance. This suggests that Pre-Annotations may standardize local boundary placement without fully standardizing higher-level segmentation granularity. Annotators may agree more closely on the timing of events they identify, while still differing in whether they split or merge adjacent activities. This reinforces the need to analyze boundary placement, event count, and semantic labels separately.

If the Assisted condition primarily reduces boundary dispersion while preserving comparable semantic-label entropy, this would support the interpretation that Pre-Annotations act as temporal scaffolds. If the Assisted condition also substantially reduces semantic-label entropy, this would provide stronger evidence of disagreement collapse or anchoring.

3.7. Anchoring and Edit-Trace Diagnostics

The central risk of Pre-Annotations is that they may anchor annotators to model priors. Consensus alignment and pairwise agreement provide indirect evidence of this risk, whereas interaction traces enable a more direct measurement. In the Assisted condition, we therefore analyze how annotators modified model-generated suggestions before submission.

Table 5. Anchoring diagnostics from Assisted-session edit traces. These metrics distinguish passive acceptance of model suggestions from active human revision.

Edit-trace diagnostic	Value
Pre-Annotation segments accepted unchanged	24.20%
Boundary edited	64.69%
Relabeled	50.12%
Split or merged	56.79%
Deleted	0.00%
Mean start-boundary displacement	1.36 s
Mean end-boundary displacement	1.09 s
Normal/abnormal label-flip rate	31.68%

We define a Pre-Annotation segment as *accepted unchanged* if its temporal boundaries and label are submitted without modification. A segment is *boundary edited* if its start or end time changes, *relabeled* if its semantic class or binary event flag changes, *split* or *merged* if its temporal structure is changed, and *deleted* if it is removed from the final annotation.

These edit categories are not mutually exclusive: a segment may be both boundary-edited and relabeled, for example. We therefore interpret the edit-trace diagnostics as evidence about revision behavior rather than as a partition of all possible edit operations.

Table 5 shows that only 24.20% of Pre-Annotation segments were accepted unchanged, while 64.69% were boundary edited, 50.12% were relabeled, and 56.79% were split or merged. The zero deletion rate indicates that annotators rarely removed model suggestions outright; instead, they primarily revised suggestions through boundary edits, relabeling, splitting, or merging. These diagnostics are necessary because higher agreement can arise through different mechanisms. If most Pre-Annotations are accepted unchanged, then the Assisted condition may reflect anchoring to model priors. If annotators frequently edit boundaries, relabel segments, split or merge suggestions, and delete incorrect proposals, then Pre-Annotations are better understood as editable scaffolds that reduce mechanical effort while preserving human oversight. Active editing does not rule out anchoring, since annotators may still make edits around model-proposed structures rather than constructing independent alternatives from scratch. The edit traces therefore rule out only a purely passive-acceptance account. They do not rule out subtler forms of anchoring, such as local adjustment around model-proposed temporal structures or selective retention of model-proposed semantic categories.

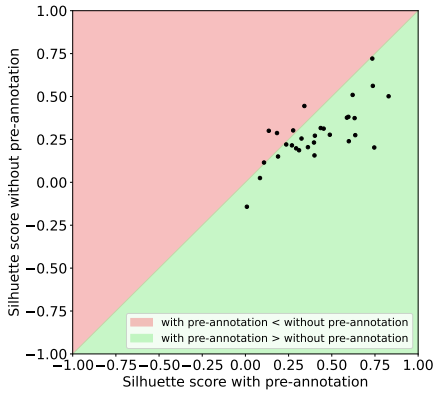


Figure 4. Latent-space standardization. Points below the diagonal indicate videos where Assisted annotations have higher Silhouette Scores than Unassisted annotations.

3.8. Latent-Space Standardization

The third audit axis measures whether annotations become more structurally coherent in the visual embedding space. We compute the Silhouette Score of human-labeled segments using CLIP image embeddings and cosine distance. Because the same encoder family is used in the Pre-Annotation engine, this metric is aligned with the model's representational geometry. We therefore interpret it as a standardization diagnostic: it measures whether assisted annotators produce boundaries that are more coherent under the embedding space used to generate suggestions.

Figure 4 shows that the Assisted condition yields higher Silhouette Scores in 25 of 30 videos. The mean score increases from 0.28 in the Unassisted condition to 0.41 in the Assisted condition, corresponding to an approximately 46.4% relative increase when computed from the rounded means. This pattern indicates that Assisted annotations are more coherent under the representational geometry used by the Pre-Annotation engine.

However, this result should be interpreted conservatively. It is a diagnostic of model-space standardization. Higher latent-space coherence is desirable when it reflects reduced arbitrary boundary jitter, but problematic when it reflects convergence toward model priors. For this reason, we interpret the Silhouette result jointly with consensus alignment, boundary-level disagreement, semantic-label entropy, and edit-trace diagnostics rather than as standalone evidence of improved annotation quality.

4. Discussion

The results reveal a distinction that is central to pluralistic data collection. Pre-Annotations increased efficiency, pairwise inter-annotator consistency, and CLIP-space standardization. Semantic-label entropy changed only slightly in aggregate, and a six-annotator human consensus diagnostic produced comparable alignment values across conditions. However, the consensus diagnostic is self-including and therefore cannot establish that assistance preserves human judgment or that anchoring is absent. The strongest supported interpretation is more modest: in this pilot setting, assistance made annotation faster and more standardized, with no large observed aggregate collapse in semantic-label entropy under the reported fixed-bin diagnostics. Higher consistency is not automatically desirable: it may indicate useful reduction of arbitrary temporal boundary jitter, or it may indicate harmful collapse of meaningful disagreement. We therefore interpret the study as an audit of how AI assistance changes the structure of annotation, not as definitive evidence that all meaningful disagreement is preserved.

The strongest empirical summary is therefore not that Pre-Annotations preserve disagreement, but that they make annotation faster and more standardized without producing a large observed change in aggregate semantic-label entropy. The remaining question is what kind of standardization assistance induces. In our results, the largest changes appear in time-on-task, pairwise agreement, boundary dispersion, and CLIP-space Silhouette Score. In contrast, semantic-label entropy changes little, and consensus alignment appears comparable under a descriptive self-including diagnostic. This pattern is more consistent with temporal scaffolding than with wholesale semantic collapse, although the evidence is not sufficient to conclude that all meaningful disagreement was preserved.

A key ambiguity is that increased agreement can be interpreted either as boundary regularization or as anchoring. Our protocol addresses this ambiguity by separating annotation time, consensus alignment, pairwise consistency, latent-space standardization, boundary variation, semantic-label entropy, and edit behavior. The central empirical signal is therefore not a single metric, but a pattern across diagnostics. In the present pilot, the strongest evidence supports efficiency gains and standardization. Stronger claims about disagreement preservation require direct evidence that semantic-label diversity, minority interpretations, and reasonable alternative segmentations remain available after assistance.

This pattern supports a practical design principle: AI assistance should be evaluated as an intervention on the distribution of human labels, not only as a way to reduce cost. A workflow can appear successful under standard metrics because it is faster and cleaner, while still being problematic if it collapses disagreement that should remain visible. Conversely, some forms of standardization may be desirable when they reduce arbitrary boundary jitter and leave annotators free to correct semantic labels.

The interaction traces are central to this distinction. Final

aggregate labels do not reveal whether annotators accepted model suggestions, edited them, rejected them, or converged independently. By preserving disaggregated traces, the audit makes it possible to analyze both the final label distribution and the process by which it was produced.

5. Limitations

This study has several limitations. First, the sample size is modest: 18 annotators, 30 videos, and six annotations per video. This design is suitable for a controlled pilot audit of one annotation workflow, but it is underpowered to rule out subtle shifts in semantic interpretation, minority-label suppression, or community-specific disagreement. A near-zero aggregate change in semantic-label entropy should therefore not be interpreted as proof that all meaningful disagreement was preserved.

Second, annotators were university-affiliated volunteers, so the results do not provide population-level evidence about broader pluralistic values. Judgments about normality, abnormality, event boundaries, and contextual relevance may vary across communities and stakeholder groups. The study evaluates how one AI-assisted workflow changes annotation behavior under controlled conditions; it does not establish how affected communities would interpret the same videos or labels.

Third, we evaluate one fixed Pre-Annotation engine and one interface. Results may differ with other models, suggestion qualities, interface designs, video domains, or annotator populations. Fourth, the Consensus Reference is an aggregate diagnostic rather than objective ground truth. Because the six-annotator reference includes the evaluated annotation, consensus-alignment values are self-including and may be optimistic. Comparable alignment therefore does not establish that anchoring is absent. Fifth, agreement and entropy metrics depend on the fixed temporal binning scheme used to discretize continuous annotations; we did not perform a bin-width sensitivity analysis in this pilot.

Finally, the task involves sensitive open-world video content. Larger studies should use paid participation, stratified recruitment, content warnings, opt-out and skip mechanisms, debriefing resources, and stakeholder-specific analysis. These results should not be used to justify deployment of surveillance or anomaly-detection systems without broader ethical review and stakeholder engagement.

6. Conclusion

We studied AI-assisted temporal video annotation as a pluralistic data-collection problem: model suggestions should be evaluated not only by whether they speed up labeling, but by whether they preserve, reshape, or collapse disagreement in ambiguous tasks.

In a controlled pilot with 18 annotators and 180 sessions, Pre-Annotations reduced annotation time for most participants and increased inter-annotator consistency. Boundary dispersion decreased, while semantic-label entropy changed only slightly in aggregate. A self-including human consensus diagnostic showed comparable alignment across conditions, but this result should be interpreted descriptively rather than as independent evidence that anchoring is absent. These results suggest that assistance acted mainly as an editable temporal scaffold in this setting. However, they do not prove that all meaningful disagreement was preserved; rather, assistance improved efficiency and standardization without an observed large aggregate collapse in semantic-label entropy under the reported diagnostics.

Our broader contribution is an audit methodology: preserve disaggregated traces, separate efficiency from disagreement collapse, avoid treating consensus as ground truth, decompose boundary and semantic disagreement, and analyze whether model suggestions are accepted or actively revised. AI-assisted annotation should be evaluated not only by whether it makes labels faster or cleaner, but by how it changes the distribution of human judgments from which datasets are built.

Impact Statement

This work studies AI-assisted data collection for high-ambiguity video annotation. Such workflows can reduce annotation effort, but they can also reshape the distribution of human judgments by anchoring annotators to model-generated suggestions. In pluralistic settings, this creates a risk that datasets become more standardized while losing disagreement that reflects meaningful differences in interpretation, context, or values.

We argue that AI-assisted annotation systems should preserve disaggregated annotator traces, document participant recruitment and task conditions, and report disagreement-related metrics rather than only aggregate labels or time savings. Such reports should also document how consensus references are constructed, whether evaluated annotations are included in those references, and how continuous annotations are discretized for metric computation. These practices can help researchers identify when AI assistance supports human judgment and when it suppresses it.

The video domains considered here may include sensitive, violent, or distressing content. Researchers and practitioners using similar workflows should provide content warnings, opt-out mechanisms, appropriate support, and fair compensation. The workflow should not be used to justify deployment of surveillance or anomaly-detection systems without broader stakeholder engagement and domain ethical review.

References

CVAT.ai Corporation. Computer vision annotation tool (cvat). https://cvat.ai/, November 2023.

Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Deng, D., Wu, J., Wang, J., Wu, Y., Xie, X., Zhou, Z., Zhang, H., Zhang, X., and Wu, Y. Eventanchor: Reducing human interactions in event annotation of racket sports videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2021.

Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., and Denton, E. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2342–2351, 2022.

Fazelpour, S. and Fleisher, W. The value of disagreement in ai design, evaluation, and alignment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 2138–2150, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3715275.3732146. URL https://doi.org/10.1145/3715275.3732146.

Feng, S., Chen, C., and Xing, Z. Video2action: Reducing human interactions in action annotation of app tutorial videos. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–15, 2023.

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J. T., Hashimoto, T., and Bernstein, M. S. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3491102.3502004. URL https://doi.org/10.1145/3491102.3502004.

Hornik, K., Feinerer, I., Kober, M., and Buchta, C. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012. doi: 10.18637/jss.v050.i10. URL https://www.jstatsoft.org/index.php/jss/article/view/v050i10.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Niebles, J. C. Dense-captioning events in videos, 2017. URL https://arxiv.org/abs/1705.00754.

Kuznetsova, A., Talati, A., Luo, Y., Simmons, K., and Ferrari, V. Efficient video annotation with visual interpolation and frame selection guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3070–3079, 2021.

Li, Y., Wang, J., Dai, X., Wang, L., Yeh, C.-C. M., Zheng, Y., Zhang, W., and Ma, K.-L. How does attention work in vision transformers? a visual analytics attempt. *IEEE Transactions on Visualization and Computer Graphics*, 29: 2888–2900, 2023. doi: 10.1109/TVCG.2023.3261935.

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017a.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts, 2017b. URL https://arxiv.org/abs/1608.03983.

Ma, J., Bai, Y., Zhong, B., Zhang, W., Yao, T., and Mei, T. Visualizing and understanding patch interactions in vision transformer. *IEEE transactions on neural networks and learning systems*, PP, 2022. doi: 10.48550/arXiv.2203.05922.

Papadopoulos, D. P., Uijlings, J. R., Keller, F., and Ferrari, V. We don't need no bounding-boxes: Training object class detectors using only human verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 854–863, 2016.

Papadopoulos, D. P., Uijlings, J. R., Keller, F., and Ferrari, V. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE international conference on computer vision*, pp. 4930–4939, 2017.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. 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.

Rasheed, H., Khattak, M. U., Maaz, M., Khan, S., and Khan, F. S. Fine-tuned clip models are efficient video learners, 2023. URL https://arxiv.org/abs/2212.03640.

Rosenberg, A. and Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical*

Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 410–420, 2007.

Shahapure, K. R. and Nicholas, C. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pp. 747–748. IEEE, 2020.

Shrestha, S., Sentosatio, W., Peng, H., Fermuller, C., and Aloimonos, Y. Feva: Fast event video annotation tool. *arXiv preprint arXiv:2301.00482*, 2023.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Sultani, W., Chen, C., and Shah, M. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Tkachenko, M., Malyuk, M., Holmanyuk, A., and Liubimov, N. Label Studio: Data labeling software. https://github.com/heartexlabs/label-studio, 2020.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

Wang, M., Xing, J., Mei, J., Liu, Y., and Jiang, Y. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2023. doi: 10.1109/TNNLS.2023.3331841.

Wu, P., Liu, j., Shi, Y., Sun, Y., Shao, F., Wu, Z., and Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision (ECCV)*, 2020.

Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., and Zhang, Y. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082, 2024.

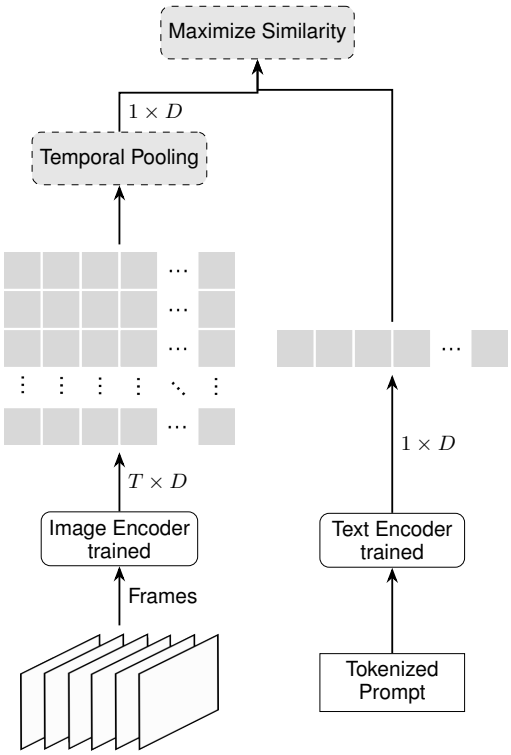


Figure 5. Cross-modal encoder used by the Pre-Annotation engine. Video frames are encoded, pooled into a video-level embedding, and aligned with text embeddings through contrastive training.

A. Fixed Pre-Annotation Engine and Model Validation

This appendix describes the Pre-Annotation engine used in the Assisted condition. The engine provides a fixed source of model-generated temporal suggestions for the annotation audit.

A.1. Video-Text Encoding

We employ a cross-modal retrieval architecture adapted for video downstream tasks (Radford et al., 2021). Given a set of video-text pairs (V, P), each video clip $V_i \in \mathbb{R}^{T\times H\times W\times C}$ is treated as a sequence of T frames. Each frame is encoded to produce frame-level features $z_i \in \mathbb{R}^{T\times D}$. A video-level embedding $v_i \in \mathbb{R}^D$ is obtained by average pooling over the T frames (Rasheed et al., 2023).

The corresponding text description p_i is processed by the text encoder to produce an embedding $t_i \in \mathbb{R}^D$ in the shared feature space. The alignment between modalities is driven by cosine similarity $\mathrm{sim}(v_i, t_i)$. We fine-tune the encoders using a contrastive objective. For a mini-batch of size B, the video-to-text loss is:

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\big(\mathrm{sim}(v_i,t_i)/\tau\big)}{\sum_{j=1}^{B}\exp\big(\mathrm{sim}(v_i,t_j)/\tau\big)}. \tag{5}$$

The total loss averages the video-to-text and text-to-video directions:

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\left(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}\right). \tag{6}$$

Figure 5 illustrates the fixed cross-modal architecture used to produce the representations for Pre-Annotation. The diagram clarifies that temporal information enters through frame-level encoding followed by pooling, while supervision is provided through video-text contrastive alignment. This supports the interpretation of the engine as a semantic proposal generator rather than a frame-perfect temporal annotator.

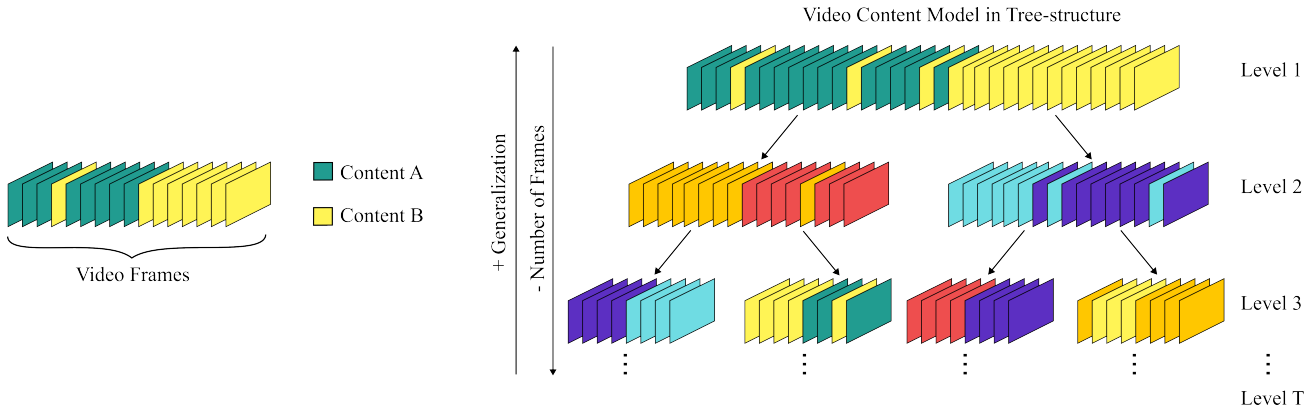


Figure 6. Hierarchical clustering of video content. Frames are recursively clustered into a tree structure using spherical k-means. Leaf nodes are converted into Pre-Annotation segments.

A.2. Hierarchical Content Modeling

To generate temporal suggestions, we model the video feature space as a hierarchical tree. This structure supports multiple levels of granularity, from broad scene changes to specific actions. We construct the tree by recursively clustering feature arrays using spherical k-means (Hornik et al., 2012). This variant maximizes cosine similarity between data points and centroids, matching the hyperspherical structure of normalized CLIP embeddings (Liang et al., 2022). The objective is:

$$\max_{r_{ij},\,\boldsymbol{\mu}_j} \sum_{i=1}^{p} \sum_{j=1}^{k} r_{ij} \, \frac{\boldsymbol{\mu}_j \cdot \mathbf{x}_i}{\|\boldsymbol{\mu}_j\| \, \|\mathbf{x}_i\|}, \tag{7}$$

subject to $r_{ij} \in \{0, 1\}$ and $\sum_{j=1}^{k} r_{ij} = 1$ for all i. Here, $\boldsymbol{\mu}_j$ is the centroid of cluster j, $\mathbf{x}_i$ is the i-th data point, and $r_{ij} = 1$ indicates assignment of $\mathbf{x}_i$ to cluster j.

The process begins with high-level clusters and recursively refines them into fine-grained nodes. We use a maximum tree depth $T_{\max} = 3$ and a semantic cohesion threshold $\tau = 0.85$, halting subdivision when within-cluster cosine similarity exceeds the threshold. Leaf nodes become Pre-Annotation segments.

A.3. Training Datasets and Setup

We trained the encoder using a composite of four datasets to expose the model to diverse visual environments, action durations, and language descriptions. UCF-Crime provides 1,900 untrimmed real-world surveillance videos totaling 128 hours (Sultani et al., 2018). UCA augments UCF-Crime with 23,542 linguistic annotations across 110.7 hours of footage (Wu et al., 2024). XD-Violence contributes 4,754 untrimmed videos totaling 217 hours (Wu et al., 2020). ActivityNet Captions adds 849 hours of footage across 20,000 videos (Krishna et al., 2017).

The cross-modal encoder was fine-tuned on an NVIDIA A100 GPU with 40 GB VRAM. We used AdamW (Loshchilov & Hutter, 2017a) with a decoupled weight decay of 1×10^{-3}. The initial learning rate was 8×10^{-6}. We used cosine learning-rate decay (Loshchilov & Hutter, 2017b):

$$\mathrm{lr}(t) = \frac{1}{2} \, \mathrm{lr}_{\mathrm{base}} \left(1 + \cos\left(\frac{t - t_{\text{warm-up}}}{T_{\text{total}} - t_{\text{warm-up}}} \, \pi \right) \right), \tag{8}$$

where T_{total} is the total number of training steps and $t_{\text{warm-up}}$ is the number of warm-up steps.

A.4. Evaluation Protocol and Results

We evaluated the encoder using video-text matching. For each text query, the model ranked the matching video clip against a candidate pool of approximately 12–13 clips, consisting of the positive clip and randomly sampled negative clips. This candidate-pool size is consistent with the random baseline in Table 6. The model predicts the most probable video-text pair by comparing embeddings with cosine similarity and normalizing scores with a softmax. During evaluation, we introduced

randomness by sampling negative clips both from within the same source video and from different videos, making the retrieval task include visually similar distractors as well as cross-video distractors. To reduce leakage, all clips derived from the same source video were assigned to the same train, validation, or test split.

We report Accuracy at Rank $\mathbb{K}$ (Acc@$\mathbb{K}$). Acc@$\mathbb{K}$ is the percentage of test samples for which the correct result appears within the top-$\mathbb{K}$ retrieved clips. Table 6 reports Acc@1 and Acc@5 for several configurations.

Table 6. Performance comparison of CLIP fine-tuning configurations.

Backbone	Batch Size	Sampled Frames	Acc@1 (%)	Acc@5 (%)
Random	–	–	8.17	39.89
ViT-B/16	10	16	**89.68**	**99.64**
	30	30	73.59	95.30
ViT-B/32	5	30	73.71	95.20
	30	5	73.71	95.20
	20	32	81.76	97.75
	100	40	59.59	86.24

The selected ViT-B/16 configuration with batch size 10 and 16 sampled frames achieved the highest Acc@1 score, 89.68%, and Acc@5 score, 99.64%. This configuration was used to generate the Pre-Annotations in the Assisted condition.

A.5. Attention Visualization

We visualized self-attention maps from the Vision Transformer to inspect whether fine-tuning shifted attention toward relevant regions in representative frames (Vaswani et al., 2017; Radford et al., 2021). We extracted multi-head attention weights, computed cumulative joint attention across layers, and isolated attention associated with the CLS token. Attention maps were averaged across heads and upscaled to the input frame resolution (Ma et al., 2022; Li et al., 2023).

Figure 7 provides a qualitative sanity check on the fine-tuned encoder. The trained attention maps appear to concentrate more on task-relevant actors, objects, or interaction regions than the baseline maps in these examples. We use this visualization only as qualitative support; the quantitative validation of the encoder is provided by the retrieval results in Table 6.

B. Annotation Interface

Figure 8 also clarifies the behavioral intervention studied in the main paper. In the Unassisted condition, annotators construct the temporal segmentation from scratch using the same workspace. In the Assisted condition, the temporal bar is pre-populated with editable suggestions, shifting the task from initial construction toward review, correction, splitting, merging, and relabeling.

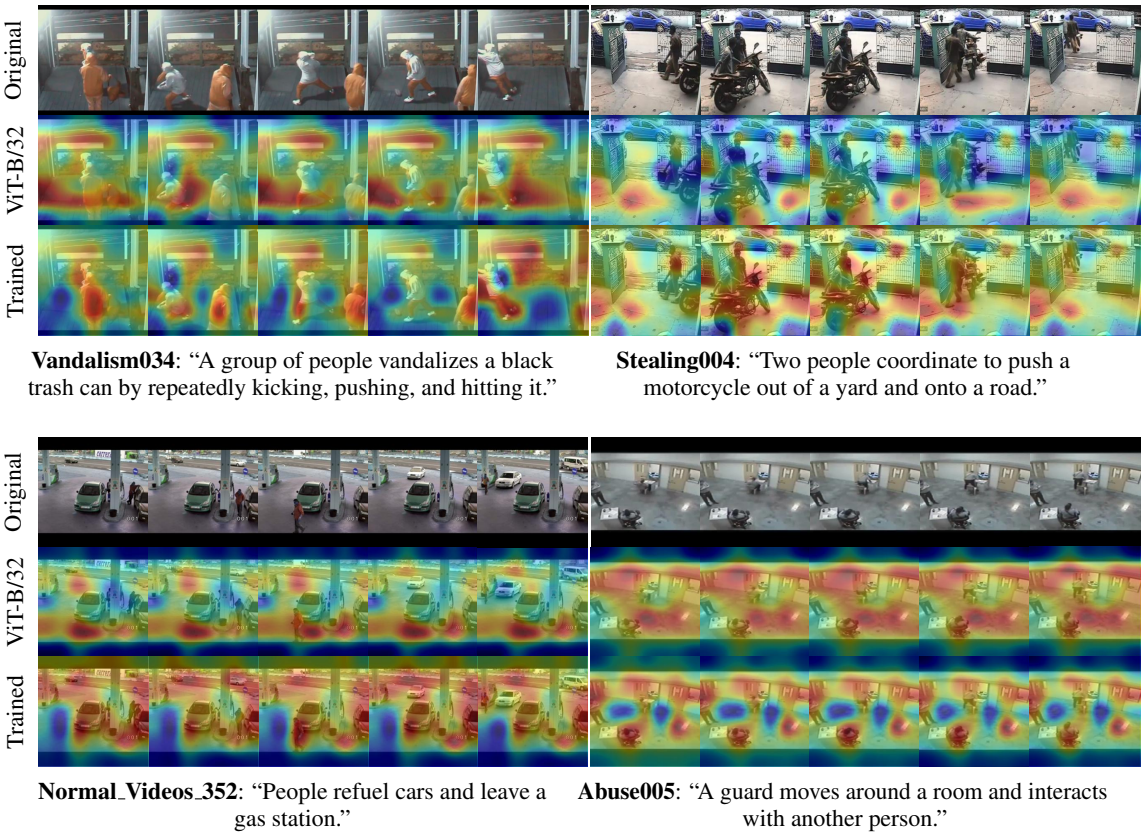
C. CrowdWorkSheets Documentation

We include a CrowdWorkSheets-style description of the annotation process (Díaz et al., 2022). The purpose is to document task subjectivity, annotator recruitment, platform choices, and known risks.

C.1. Task Formulation

At a high level, what are the subjective aspects of your task? The task is subjective in two ways. First, annotators decide how to group frames into temporally coherent events. The boundary between the end of one event and the beginning of the next can be gradual. Second, annotators classify each segment as normal or abnormal. The guide defined abnormal events as criminal, accidental, or catastrophic incidents, but applying this definition still requires contextual interpretation.

What assumptions do you make about annotators? We assume that annotators follow the instructions, watch each video before annotating, and use their own judgment when boundaries or event categories are ambiguous. We also assume that disagreements are expected and informative in this task.



Vandalism034: "A group of people vandalizes a black trash can by repeatedly kicking, pushing, and hitting it."

Stealing004: "Two people coordinate to push a motorcycle out of a yard and onto a road."

Normal_Videos_352: "People refuel cars and leave a gas station."

Abuse005: "A guard moves around a room and interacts with another person."

Figure 7. Attention map visualization on representative UCF-Crime samples. Each merged image contains three stacked rows: original frame, baseline ViT-B/32 attention, and fine-tuned attention.

How did you choose the wording of the task instructions? The instructions were designed to give annotators a concrete operational definition while leaving room for judgment. The abnormal-event definition was intentionally broad enough to cover the relevant video categories without forcing frame-perfect agreement. The guide was reviewed by multiple researchers and updated after early annotation batches.

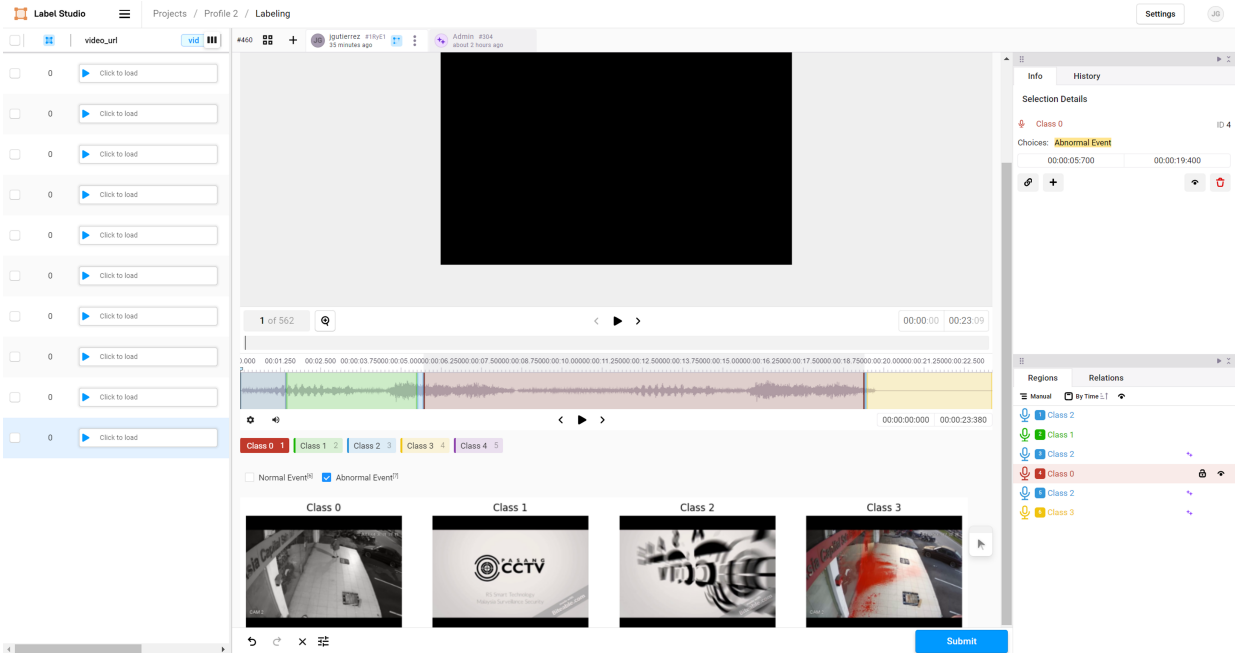
What risks did the task pose for annotators? The task involved videos that could include violent, explicit, catastrophic, or otherwise distressing content. Annotators were informed of the content type before participation. Videos were filtered before the study to remove material judged too intense for the annotation setting.

What precise instructions were provided to annotators? Annotators were asked to annotate open-world videos. Each annotator labeled 10 videos: 5 without Pre-Annotations and 5 with model-generated Pre-Annotations. For each video, annotators grouped frames with similar content by marking temporal segments and assigning class labels. They also marked each segment as normal or abnormal. Annotators were instructed to watch the full video before beginning annotation and to use the guide when applying the normal/abnormal distinction.

C.2. Selecting Annotators

Were certain perspectives privileged? The study used a controlled university-affiliated cohort. The aim was to measure how the annotation workflow changes behavior under a consistent study protocol.

Were sociodemographic characteristics used to select annotators? Annotators were volunteers affiliated with a university institution. They had at least university-level academic background and no prior exposure to the specific test clips. The cohort was balanced in gender and included ages ranging from 18 to 50.



(a) Main annotation workspace with video player and class controls.

(b) Temporal timeline with suggested segments and binary event controls.

Figure 8. Customized Label Studio interface. In the Assisted condition, Pre-Annotations appear on the temporal bar as suggestions that annotators can copy, edit, split, merge, delete, or relabel.

Could annotator characteristics have affected the annotations? Yes. Judgments about normality, abnormality, event boundaries, and contextual relevance can vary across communities and backgrounds. This is why the main paper treats the study as a controlled annotation audit and does not interpret the Consensus Reference as objective ground truth.

Are affected communities represented in the annotator pool? The annotator pool is not designed to represent all communities that may be affected by video analysis systems. The study evaluates the annotation workflow and preserves disaggregated traces for analysis. Broader deployment would require stakeholder-specific recruitment and analysis.

C.3. Platform and Infrastructure Choices

What annotation platform did you use? We used Label Studio (Tkachenko et al., 2020).

Why was this platform chosen? Label Studio is open source, configurable, and supports temporal video annotation. It also records timing information and allows integration with model-generated Pre-Annotations.

Did the platform meet the requirements of the task? Yes. The customized interface supported video playback, temporal segmentation, segment labeling, binary event flags, and Pre-Annotation review.

What communication channels were available? Annotators could ask clarification questions directly or through email. Questions about tool usage were answered. Questions about ambiguous content were answered by referring annotators back

to the written guide and the principle that reasonable disagreement was expected.

How were annotators compensated? Annotators participated voluntarily and were not financially compensated.

C.4. Dataset Analysis and Evaluation

How do you define annotation quality in this context? Quality is not defined as agreement with a single ground truth. We evaluate the annotation process through multiple diagnostics: annotation time, alignment with an aggregate Consensus Reference, pairwise inter-annotator consistency, and latent-space standardization.

Did you analyze disagreement patterns? Yes. Section 3 analyzes disagreement through pairwise inter-annotator consistency and compares Assisted and Unassisted conditions. It also measures latent-space standardization to distinguish boundary regularization from simple agreement increase.

How do individual annotator responses relate to released labels? The released artifact preserves annotator-level traces rather than only a final aggregated label. This allows future work to study disagreement, revision behavior, and the effect of model suggestions on the annotation process.

Ethics and participant protection. Participants were informed before the study that the videos could contain violent, abnormal, catastrophic, or otherwise distressing content. Participation was voluntary, and participants could stop the task at any time. The study used a controlled university-affiliated cohort and filtered videos before annotation to remove material judged too intense for the study setting. Because unpaid annotation of distressing content raises ethical concerns, we treat the volunteer design as a limitation of the pilot and recommend fair compensation, explicit opt-out mechanisms, and debriefing resources for larger-scale studies.