AIRRELCOMP: A DIAGNOSTIC BENCHMARK FOR VIDEO-TEXT COMPOSITIONAL UNDERSTANDING IN AIRPORT SURVEILLANCE

Anonymous authors

ABSTRACT

Understanding continuous multi-entity airport activities for safety assessment requires resolving temporal order, spatial configuration, and entity participation beyond merely recognizing entities and actions. Broad video-text matching can conceal errors in these relations. We introduce AirRelComp, a diagnostic benchmark for fine-grained compositional understanding in airport surface videos. AirRelComp organizes Temporal, Spatial, and Entity Relations under a common event schema. We manually annotate positive video-text pairs from surveillance videos and construct four types of controlled negatives using predefined rules. Each negative changes one designated relation while preserving the remaining event structure. AirRelComp contains nearly 4K controlled comparisons and supports similarity comparison for video-text models and binary choice for large multi-modal models (LMMs). Evaluations reveal limited zero-shot transfer in video-text models and uneven discrimination across disruption types in LMMs. We therefore develop a learning framework that integrates supervision from positive–negative correspondence and relation identity with representative visual patch selection. With the same backbone, the framework raises Overall Accuracy from 63.45% under positive-only fine-tuning to 78.98%. Ablations show that the structured use of controlled negatives outperforms generic negative augmentation. LoRA fine-tuning on AirRelComp further substantially improves LMM binary-choice accuracy. These results show that AirRelComp diagnoses compositional weaknesses and provides effective supervision for improving relation discrimination.

1 INTRODUCTION

Airport surface surveillance continuously records operations in which aircraft, vehicles, and ground crew members coexist and act concurrently. Understanding these activities requires modeling the temporal unfolding of an activity, the locations and movements of entities, and the entities participating in each event, rather than recognizing entities and actions in isolation Sadhu et al. (2021). These entities can also differ substantially in visual scale and spatial distribution, leaving relevant activity evidence dispersed across the scene. Inspired by the Event-Indexing Model Zwaan & Radvansky (1998), we organize these aspects into three relational dimensions: Temporal, Spatial, and Entity Relations. Large-scale pretraining has broadened the understanding capabilities of video-language models and large multi-modal models (LMMs) Xue et al. (2023); Wang et al. (2024b); OpenAI (2026c). However, broad video-text alignment can be established through overall semantic correspondence without requiring sensitivity to fine-grained compositional relations. Such insensitivity remains difficult to diagnose through conventional video-language evaluations. Recent benchmarks have therefore moved toward finer-grained temporal, compositional, and interaction understanding Du et al. (2024); Kim et al. (2025); Saravanan et al. (2025); Aparcedo et al. (2026). However, in continuous multi-entity events, temporal evolution, spatial configuration, and entity participation are jointly expressed within the same event description, as illustrated in Fig. 1. Existing evaluations often examine these aspects separately or alter several parts of the event at once, making it difficult to isolate a model's sensitivity to a specific relation while preserving the remaining event structure.

We introduce AirRelComp, a diagnostic benchmark and data resource for learning fine-grained compositional relations in airport surface videos. Human annotation first produces temporally aligned factual video–text pairs, from which we construct controlled compositional negatives along Tem-

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.

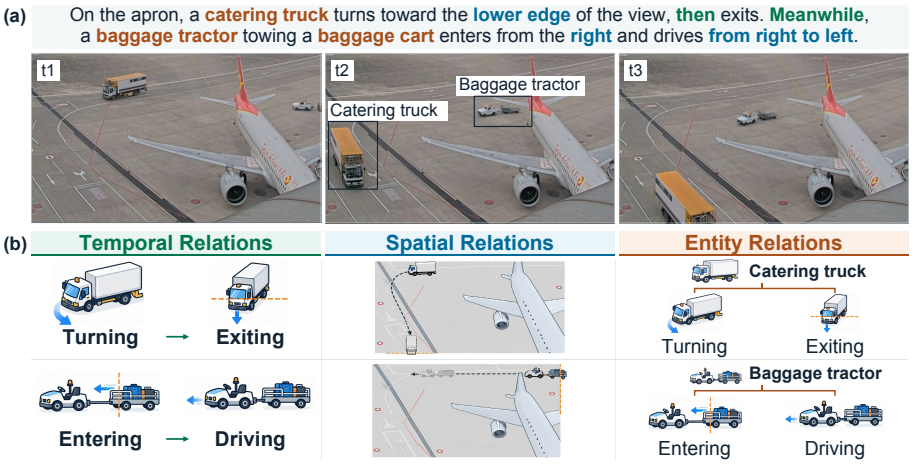


Figure 1: An example of concurrent airport activities. (a) Selected surveillance frames and the corresponding positive text. (b) Schematic illustrations of Temporal, Spatial, and Entity Relations jointly expressed in the same event description. Directions refer to the camera view.

poral, Spatial, and Entity Relations, covering four disruption types: *Spatial Location Replacement*, *Spatial Direction Reversal*, *Temporal Stage Reordering*, and *Entity Participation Swapping*. Each negative alters only the designated relation while preserving the remaining event structure. Predefined rules determine the relation change, while GPT-5.5 OpenAI (2026b) realizes it in natural language. AirRelComp further defines two benchmark tasks for video–text models and LMMs.

The controlled comparisons in AirRelComp also provide structured supervision through positive–negative correspondence and the identity of the changed relation. We therefore introduce RelCompLoss, which combines negative preference for distinguishing a positive text from its controlled negatives with relation preference over the corresponding relation representations. We further introduce SpaSelection to reduce visual redundancy while retaining representative patches from multiple regions. Experiments show that large-scale external pretraining provides only limited zero-shot discrimination on AirRelComp. Using CLIP-ViP Xue et al. (2023) as the backbone, our full framework raises Overall Acc. from 63.45% under positive-only fine-tuning to 78.98% and Comp. Score from 14.54% to 35.78%. LMMs also exhibit uneven performance across the disruption types, while LoRA Hu et al. (2021) fine-tuning Qwen3.5-VL-9B on AirRelComp improves Overall Acc. from 74.15% to 89.82%, showing that the controlled comparisons provide effective supervision for both model classes. Our main contributions are summarized as follows:

- We introduce AirRelComp, a diagnostic benchmark that explicitly organizes continuous multi-entity airport activities under a common event schema for Temporal, Spatial, and Entity Relations. It evaluates relation discrimination through four types of controlled disruption for video–text models and LMMs.
- We propose a learning framework that exploits the supervision provided by AirRelComp. RelCompLoss learns positive-negative and relation preferences from controlled negatives, while SpaSelection retains representative visual patches under multi-entity interference.
- Evaluations of pretrained and fine-tuned video–text models and open- and closed-source LMMs reveal persistent weaknesses in fine-grained relation discrimination. Learning from AirRelComp shows the benefit of controlled relation supervision, and our framework achieves higher video–text retrieval recall on the validation split.

2 Related Work

2.1 Airport Surveillance Video Understanding

Airport video surveillance provides direct observations of aircraft, vehicles, and personnel for operational analysis and safety assessment. Research on airport video encompasses object perception, motion analysis, and the interpretation of surface activities Zhao et al. (2026); Zhou et al. (2024). Tracking and foreground segmentation recover object locations and distinguish moving from station-

ary aircraft Zhang et al. (2024b). Beyond object perception, prior work addresses servicing activity recognition Fusier et al. (2007), aircraft motion and push-back time estimation Thai et al. (2022), and safety-related conflict detection Zhang et al. (2024a). Existing studies exploit spatial, temporal, and entity cues for specific perception or operational tasks, leaving their organization within a continuous multi-entity event largely implicit. AirRelComp instead organizes these relations under a common event schema and evaluates them through controlled video–text comparisons.

2.2 Video Compositionality and Diagnostic Benchmarks

Compositional video understanding requires models to move beyond coarse video–text alignment and capture fine-grained relations in dynamic visual content. Image–text compositional benchmarks such as Winoground Thrush et al. (2022) examine whether models can correctly match images with texts containing the same words in different orders. However, these static evaluations cannot capture the dynamic nature of videos, where objects interact and events evolve over time. Video-language benchmarks have expanded beyond broad video–text alignment evaluation Anne Hendricks et al. (2017); Krishna et al. (2017). Recent benchmarks examine temporal understanding Li et al. (2024); Kesen et al. (2024); Du et al. (2024); Liu et al. (2024) and broader compositional and relational semantics Kim et al. (2025); Bansal et al. (2024); Saravanan et al. (2025); Aparcedo et al. (2026); Fu et al. (2025); Zhu et al. (2026). Although these benchmarks advance fine-grained video diagnosis, they typically target temporal, compositional, or interaction phenomena rather than organizing multiple relations within a shared event representation. AirRelComp addresses this gap by modeling Temporal, Spatial, and Entity Relations within the same event schema and constructing matched comparisons that change one relation at a time while preserving the remaining event structure.

2.3 Video-Text Alignment and Compositional Learning

Effective video–text alignment requires both retaining informative visual evidence and learning its fine-grained correspondence with language. In multi-entity videos, informative cues can be sparsely distributed across the scene and obscured by redundant visual content, motivating selective video representations that retain relevant spatiotemporal evidence while reducing redundancy Wei et al. (2022); Li et al. (2023); Zhuang et al. (2026). For learning strategies, contrastive alignment has become a prevailing paradigm, from CLIP Radford et al. (2021) to video-language models with temporal modeling and large-scale video–text pretraining Xue et al. (2023); Wang et al. (2024b). However, ordinary mismatched pairs provide limited supervision for distinguishing a positive text from a closely related alternative. Compositional learning therefore introduces harder or controlled negatives into contrastive training Yuksekgonul et al. (2023); Du et al. (2024), while later methods further exploit them through entailment or explicit preference supervision Bansal et al. (2024); Kim et al. (2025). Although these studies strengthen fine-grained video–text learning, they do not jointly exploit the correspondence between each positive and its controlled negatives and the identity of the changed relation. SpaSelection preserves representative visual evidence across multiple regions, while RelCompLoss uses both signals to learn negative and relation preferences.

3 AirRelComp Benchmark

3.1 Benchmark Construction

We construct AirRelComp from 118 continuous surveillance videos totaling 15.19 hours, recorded from two fixed viewpoints at a civil airport. Eight undergraduate students majoring in engineering annotate the videos to produce temporally aligned positive video–text pairs (v_i, t_i^+). Our construction follows two principles. First, the positive and negative texts retain the same words whenever possible and differ only in the designated relation, thereby limiting the usefulness of bag-of-words cues. Second, an event schema informed by the Event-Indexing Model Zwaan & Radvansky (1998) organizes each positive text into event content and three relational dimensions. **Temporal Relations** capture activity ordering, transitions, and concurrency; **Spatial Relations** capture entity locations and movement; and **Entity Relations** capture participation assignments between entities and event units. Further source-video statistics and annotation details are provided in Appendix A.1.

Based on this event schema, we construct four types of controlled disruptions by altering one relation at a time. For each positive video-text pair (v_i, t_i^+), we construct K_i controlled negative texts, where $1 \leq K_i \leq 4$. Each negative text changes only one target relation while leaving the others unchanged,

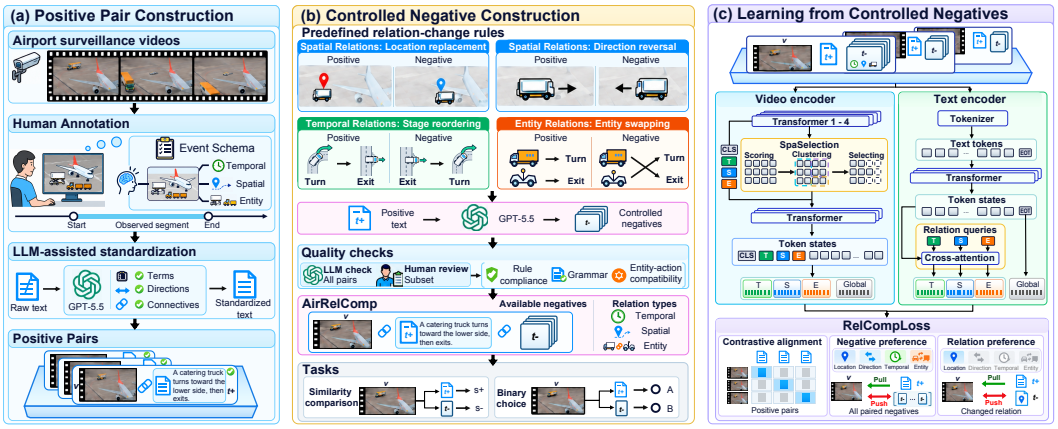


Figure 2: Overview of AirRelComp construction and learning from controlled negatives. (a) Human annotation and LLM-assisted standardization produce positive video-text pairs. (b) Four predefined relation-change rules and quality checks yield controlled negatives for similarity comparison and binary choice. (c) On the training split, SpaSelection retains representative visual patches, while RelCompLoss learns from positive-negative correspondence and disruption types.

yielding a semantically plausible counterfactual to the depicted video. AirRelComp is organized as:

$$\mathcal{D} = \left\{ \left(v_i, t_i^+, \mathcal{N}_i \right) \right\}_{i=1}^{N}, \qquad 1 \le |\mathcal{N}_i| \le 4, \tag{1}$$

where $\mathcal{N}_i$ contains the controlled negative texts and their corresponding disruption labels for the i-th positive pair. Fig. 2 summarizes the construction of AirRelComp.

3.1.1 Positive Video-Text Pair Construction

We construct positive video–text pairs from surveillance videos through human annotation and LLM-assisted standardization, as illustrated in Fig. 2(a). Following cognitive theory that represents events along multiple situational dimensions, including time, space, and participating entities Zwaan & Radvansky (1998), annotators track aircraft, vehicles, and personnel across space and time, recording visually observable entities and actions together with their locations, motion directions, temporal stage transitions, and participation. For each annotated activity span, annotators mark its start and end timestamps and write a positive text t_i^+ for the corresponding video segment v_i. The detailed annotation protocol and formal event representation are provided in Appendix A.1.

3.1.2 Controlled Compositional Negative Construction

We derive each negative directly from its positive text under predefined relation-change rules, using minimal wording changes to limit bag-of-words shortcuts. Fig. 2 (b) illustrates four controlled compositional disruptions. The first two address complementary aspects of Spatial Relations, while the remaining two target Temporal and Entity Relations, respectively. 1) **Spatial Location Replacement (Loc-Replace)** changes the reported location of an entity or activity, including its appearance, entry, exit, or stopping position, while preserving the underlying activity and event order. It evaluates sensitivity to spatial location beyond activity recognition. 2) **Spatial Direction Reversal (Dir-Reverse)** reverses motion direction, exchanges path endpoints, or applies a predefined motion target substitution while preserving the motion event and temporal order. It evaluates sensitivity to movement direction beyond motion detection. 3) **Temporal Stage Reordering (Stage-Reorder)** reorders complete activity stages performed by the same entity while preserving the content of each stage. It evaluates sensitivity to temporal stage order under nearly unchanged vocabulary. 4) **Entity Participation Swapping (Entity-Swap)** reassigns existing entities to existing activities while preserving the spatiotemporal content of each activity and the overall temporal organization. It evaluates sensitivity to the association between entities and activities beyond their separate recognition.

We design one prompt template for each of the four controlled compositional disruptions and use ChatGPT-5.5 OpenAI (2026b) to express the change prescribed by each template. Each template specifies the permitted substitutions or stage reordering and requires all non-target event content

Table 1: Pair counts and lexical overlap by disruption type.

Metric	Loc-Replace	Dir-Reverse	Stage-Reorder	Entity-Swap	Overall
Pairs	1,414	1,224	755	462	3,855
Jaccard (%)	93.32	97.50	99.86	94.24	96.04
Retention (%)	96.57	98.77	99.93	96.72	97.95

and wording to remain unchanged, except for necessary grammatical adjustments and compatibility edits permitted by the rule. Appendix A.2 describes the prompt constraints and generation process.

3.1.3 Quality Assurance

We screen every generated pair for rule compliance, grammar, and entity–action compatibility using a separate text-only LLM checking prompt. Errors identified during human review guide targeted LLM corrections and searches for similar errors in other pairs. Across all 3,855 pairs in Table 1, Token Jaccard and positive vocabulary retention average 96.04% and 97.95%, respectively, supporting high lexical overlap. Appendix A.3 describes the quality checks and lexical control metrics.

3.2 Benchmark Tasks Definition

We define two benchmark tasks for AirRelComp: similarity comparison for video–text models and binary choice for LMMs. Following established compositional evaluation Yuksekgonul et al. (2023); Kim et al. (2025), both tasks use the same controlled comparison triplet $(v_i, t_i^+, t_{i,k}^-)$, consisting of a video, its positive text, and one controlled negative text. For each instance, the model must prefer the positive text over the controlled negative. We report accuracy for each disruption type and across all instances, and compute the Comprehensive Score from the four category accuracies.

Similarity Comparison for Video-Text Models. For video–text models, we compute the cross-modal similarity between the video and each candidate text. A prediction is correct when the positive text receives the higher score. Ties are counted as incorrect.

Binary Choice for Large Multi-modal Models (LMMs). For large multi-modal models, we adapt the same comparison to their generative output format. Each model receives a sequence of sampled video frames and the two candidate texts. A task-specific prompt introduces the airport setting, defines the comparison task, and asks the model to select the text that best matches the video. The response is restricted to the corresponding option only, without additional explanation.

4 Method

We use CLIP-ViP Xue et al. (2023) as the base model for efficient and scalable video and text alignment. We retain the standard bidirectional contrastive objective as the base learning objective. However, dense visual tokens can weaken the local activity and motion cues needed for fine-grained relation understanding, and standard contrastive learning does not identify the relation responsible for a mismatch. We therefore extend CLIP-ViP at the representation and learning levels, as illustrated in Fig. 2(c). Spatiotemporal token selection preserves local activity regions and motion changes. To learn from controlled relation negatives, we group each video, its positive text, and its corresponding AirRelComp negatives in the same batch for global and relation supervision.

4.1 Sparse Spatiotemporal Token Selection

Airport videos contain large areas of persistent background, while local activities and motion changes occupy only a small portion of the visual sequence. We therefore introduce SpaSelection after the fourth CLIP-ViP Transformer block, where patch tokens retain CLIP-ViP temporal encoding. It scores these tokens, uses frame-wise clustering to preserve spatial coverage, and retains K representative patches per frame while keeping the original `CLS` token and three video proxies.

4.1.1 Per-frame Spatiotemporal Patch Scoring

Each patch receives three scores from the fourth Transformer block. Let $\widehat{\mathbf{x}}_{t,n}$ denote the ℓ_2-normalized feature of patch n in frame t, and let $A^h_{m,t,n}$ denote the attention from special token

m to this patch in head h. Proxy attention averages the attention received from the special tokens. Feature change measures variation at the same patch position across adjacent frames, while local displacement captures nearby movement:

$$s_{t,n}^{\text{proxy}} = \frac{1}{HM}\sum_{h=1}^{H}\sum_{m=1}^{M} A_{m,t,n}^{h}, \qquad s_{t,n}^{\text{change}} = 1 - \langle \widehat{\mathbf{x}}_{t,n}, \widehat{\mathbf{x}}_{t-1,n} \rangle ,$$
$$j_{t,n}^{\star} = \arg\max_{j \in \mathcal{N}_r(n)} \langle \widehat{\mathbf{x}}_{t,n}, \widehat{\mathbf{x}}_{t-1,j} \rangle , \qquad s_{t,n}^{\text{disp}} = \left[\left\langle \widehat{\mathbf{x}}_{t,n}, \widehat{\mathbf{x}}_{t-1,j_{t,n}^{\star}} \right\rangle \right]_{+} \frac{\left\| \mathbf{p}_n - \mathbf{p}_{j_{t,n}^{\star}} \right\|_2}{\sqrt{2}\, r}, \tag{2}$$

where $[\cdot]_+$ denotes the positive part, H and M denote the numbers of attention heads and special tokens used for scoring, respectively, $\mathbf{p}_n$ denotes the patch grid coordinate, and $\mathcal{N}_r(n)$ contains positions within radius r in the previous frame. Each score is independently normalized over the P patches in every frame using min-max normalization $\mathcal{N}_t(\cdot)$. The final score $s_{t,n}$ is computed as:

$$s_{t,n} = \mathcal{N}_t\big(s_{t,n}^{\text{proxy}}\big) + 0.5\,\mathcal{N}_t\Big(s_{t,n}^{\text{change}}\Big) + 0.5\,\mathcal{N}_t\Big(s_{t,n}^{\text{disp}}\Big). \tag{3}$$

4.1.2 Intra-frame Clustering for Multi-entity Coverage

In AirRelComp videos, an aircraft often occupies much of the frame, while vehicles and personnel appear in smaller, dispersed regions. Direct Top-K selection using the scores in Eq. 3 may therefore retain several similar patches from the aircraft region and overlook smaller entities and their activities. We therefore cluster the normalized patch features independently within each frame using deterministic cosine K-means with C clusters, where $C \le K$. We retain the highest scoring patch from each non-empty cluster and fill the remaining positions in descending score order until K patches are selected. Deterministic clustering and selection details are provided in Appendix B.4.

Let $\mathcal{S}_t$ denote the selected patch indices in frame t. The selector produces

$$\mathbf{Z}^{\text{sel}} = \text{Concat}\left(\mathbf{Z}^{\text{special}}, \{\mathbf{x}_{t,n} \mid n \in \mathcal{S}_t\}_{t=1}^{T}\right), \quad |\mathcal{S}_t| = K, \tag{4}$$

where $\mathbf{Z}^{\text{special}}$ contains the original CLIP `CLS` token and three video proxies. The selected patches are concatenated in temporal order and jointly processed by the remaining CLIP-ViP blocks, where patch-token attention is frame-wise and the video proxies aggregate across frames.

4.2 Relation Learning from Controlled Negatives

AirRelComp exposes compositional differences through spatial, temporal, and entity relations. To strengthen the compositional understanding of CLIP-ViP, we extend the standard contrastive objective with a relation preference loss and a negative preference loss. We term the resulting objective RelCompLoss. RelCompLoss encourages the model to assign higher similarity to positive texts and associate each controlled negative with its corresponding relation dimension. Our framework starts with the standard bidirectional InfoNCE objective, which distinguishes each matched video-text pair from other positive pairs in the mini-batch. Controlled negatives remain outside this base InfoNCE.

Following the three relation dimensions, we introduce three learned queries after the text Transformer. They attend to the final text token states through a shared cross attention module to produce spatial, temporal, and entity representations. We assign the three video proxies inherited from CLIP-ViP Xue et al. (2023) to the same roles, forming corresponding video and text relation branches. The disruption type determines the branch supervised by the relation preference loss. Loc-Replace and Dir-Reverse use the spatial branch S, Stage-Reorder uses the temporal branch T, and Entity-Swap uses the entity branch E. For the k-th controlled negative associated with video v_i, let $r_{i,k} \in \{S, T, E\}$ denote its assigned branch. The cosine similarities of the positive and negative pairs on this branch are denoted by $s_i^{r_{i,k},+}$ and $s_{i,k}^{r_{i,k},-}$, respectively. The loss $\mathcal{L}_{\text{Rel}}$ is defined as:

$$\mathcal{L}_{\text{Rel}} = \frac{1}{N_{\text{neg}}}\sum_{i=1}^{B}\sum_{k=1}^{K_i} \text{softplus}\left(\frac{s_{i,k}^{r_{i,k},-} - s_i^{r_{i,k},+}}{\tau_r}\right), \tag{5}$$

where K_i is the number of applicable controlled negatives for the i-th positive pair, $N_{\text{neg}} = \sum_{i=1}^{B} K_i$, and τ_r is the relation temperature. Appendix B.2 details the text relation module.

We further use the controlled negatives to learn the preference between the positive text and its corresponding controlled negatives under the same video context. Extending preference learning over controlled text changes Kim et al. (2025), we place each positive video-text pair and all its applicable controlled negatives in the same mini-batch. This organization preserves their correspondence and allows multiple relation changes for the same video to contribute to one update. For each modality, we separately ℓ_2-normalize the global, spatial, temporal, and entity embeddings. We then scale each relation embedding by $\sqrt{\rho/3}$, concatenate the four components, and normalize the resulting representation again. Both $\mathcal{L}_{\mathrm{InfoNCE}}$ and $\mathcal{L}_{\mathrm{Neg}}$ use the cosine similarity $s(v,t)$ between these fused video and text representations. For the paired comparisons, we define $s_i^+ = s(v_i, t_i^+)$ and $s_{i,k}^- = s(v_i, t_{i,k}^-)$. The controlled negatives can differ in both number and difficulty. We therefore aggregate their pairwise margins using a K_i-normalized log-mean-exp. This aggregation retains every applicable negative while placing greater optimization weight on negatives with higher video similarity. The normalization prevents positive pairs with more negatives from dominating training. The negative preference loss is defined as:

$$\mathcal{L}_{\mathrm{Neg}} = \frac{1}{B}\sum_{i=1}^{B}\log\left[1 + \frac{1}{K_i}\sum_{k=1}^{K_i}\exp\left(\gamma\left(s_{i,k}^- - s_i^+\right)\right)\right], \tag{6}$$

where γ is the detached exponentiated CLIP logit scale.

The final RelCompLoss is formulated as:

$$\mathcal{L}_{\mathrm{RelComp}} = \mathcal{L}_{\mathrm{InfoNCE}} + \alpha\,\mathcal{L}_{\mathrm{Neg}} + \beta\,\mathcal{L}_{\mathrm{Rel}}, \tag{7}$$

where α and β control the contributions of $\mathcal{L}_{\mathrm{Neg}}$ and $\mathcal{L}_{\mathrm{Rel}}$, respectively. At inference, we use the same fused representations without requiring disruption labels.

5 Experiments

In this section, we evaluate existing models on AirRelComp and investigate its capability for diagnosing and learning fine-grained spatiotemporal compositional relations. Appendix C.1 examines textual artifacts and language bias, while the remaining sections of Appendix C analyze parameter sensitivity, video–text retrieval, and qualitative results.

5.1 Experimental Setup

We split AirRelComp by source video into 1,081/154/309 positive video-text pairs and 2,705/384/766 controlled comparisons for training, validation, and testing, respectively. The 766 test comparisons are further manually verified after the split to ensure evaluation reliability. Zero-shot video-text baselines use their released checkpoints without AirRelComp training, while fine-tuned baselines are trained only on the 1,081 positive training pairs. RelCompLoss and Qwen3.5-VL-9B adaptation additionally use the 2,705 training comparisons. All models are evaluated on the same 766 test comparisons. We report accuracy for each of the four disruption types and Overall Accuracy (Overall Acc.) over all controlled comparisons. Following Kim et al. (2025), we also report the Comprehensive Score (Comp. Score) to measure balanced performance across the four disruption types. Ties in LMM responses count as incorrect. Video–text checkpoints are selected by retrieval performance on the 154 positive validation pairs, while the adapted LMM is selected by Overall Acc. on the 384 validation comparisons. Detailed metric definitions are provided in Appendix B.1.

We initialize our framework from the released CLIP-ViP-B/32 checkpoint Xue et al. (2023) and use 64 frames per video. Video–text baselines follow their released input configurations. All LMMs receive the same cached sequence of 64 uniformly sampled frames and share the prompt, A/B assignment protocol, and deterministic decoding. We adapt Qwen3.5-VL-9B with LoRA while keeping its vision encoder frozen. Full training and adaptation details are provided in Appendix B.2.

5.2 Main Results

We evaluate AirRelComp under the two benchmark tasks, including similarity comparison with pretrained and fine-tuned video–text models and binary choice with open- and closed-source LMMs. Baseline pretraining, initialization, and input settings are provided in Appendix B.3.

Table 2: Similarity comparison of video-text models on the AirRelComp benchmark.

Method	Setting	Loc-Replace	Dir-Reverse	Stage-Reorder	Entity-Swap	Overall Acc.	Comp. Score
Large-scale Pretrained Models:							
InternVideo2 Wang et al. (2024b)	Zero-shot	41.84	45.31	41.78	50.54	43.99	4.00
ViFi-CLIP Rasheed et al. (2023)		58.51	48.16	45.89	59.14	52.87	7.65
ViCLIP Wang et al. (2024a)		39.72	36.33	48.63	52.69	41.91	3.70
LanguageBind-Video Lin et al. (2024)		50.71	51.43	61.64	55.91	53.66	8.99
CLIP-ViP Xue et al. (2023)		42.91	39.18	42.47	55.91	43.21	3.99
General Initialized Models:							
CLIP4Clip Luo et al. (2022)	Fine-tuned	60.28	50.61	38.36	69.89	54.18	8.18
NeighborRetr Lin et al. (2025)		57.80	58.78	40.41	59.14	54.96	8.12
TV-ProxyNet Zhang et al. (2025)		74.11	61.22	56.16	60.22	64.88	15.34
GLSCL Yang et al. (2026)		68.79	57.96	46.58	63.44	60.44	11.78
CLIP-ViP Xue et al. (2023)		68.79	63.27	55.48	60.22	63.45	14.54
Our Framework:							
CLIP-ViP Xue et al. (2023)	**RelCompLoss**	84.04	71.42	**79.45**	61.29	76.37	29.23
CLIP-ViP Xue et al. (2023)	**SpaSelection + RelCompLoss**	**84.75**	**75.51**	78.77	**70.97**	**78.98**	**35.78**

Table 3: Binary choice evaluation of large multi-modal models on the AirRelComp benchmark.

Method	Setting	Loc-Replace	Dir-Reverse	Stage-Reorder	Entity-Swap	Overall Acc.	Comp. Score
Open-source Models:							
Qwen2.5-VL-7B Hui et al. (2024)	Zero-shot	60.28	46.53	59.59	61.29	55.87	10.24
Qwen3-VL-8B Bai et al. (2025)		60.99	65.71	93.15	77.42	70.63	28.91
Qwen3.5-VL-4B Team (2026)		66.67	60.00	85.62	76.34	69.32	26.15
Qwen3.5-VL-9B Team (2026)		68.09	69.80	92.47	75.27	74.15	33.07
InternVL3-8B Zhu et al. (2025)		60.28	58.37	82.19	77.42	65.93	22.39
VideoChat-R1.5 Yan et al. (2026)		60.64	60.41	78.77	75.27	65.80	21.72
Owl-Con Bansal et al. (2024)		47.17	51.88	72.45	43.93	53.23	7.79
Closed-source Models:							
Gemini 3.5 Flash Kavukcuoglu et al. (2026)	Zero-shot	90.07	**85.71**	81.50	79.57	85.77	50.07
ChatGPT-5.4 mini OpenAI (2026a)		73.40	62.86	85.62	68.82	71.80	27.19
ChatGPT-5.6 luna OpenAI (2026c)		83.33	78.78	89.04	**84.95**	83.16	49.65
ChatGPT-5.6 sol OpenAI (2026c)		84.75	84.08	94.52	82.80	86.17	55.77
Adapted Models:							
Qwen3.5-VL-9B + LoRA	Fine-tuned	**93.62**	84.90	**95.89**	81.72	**89.82**	**62.28**

5.2.1 Evaluating Video-Text Models with Similarity Comparison

Table 2 reports results for zero-shot transfer, positive-only fine-tuning, and our framework trained with controlled negatives. The pretrained models exhibit limited zero-shot transfer to AirRelComp. Despite training on external datasets, their Overall Accuracy ranges from 41.91% to 53.66%, close to the random baseline, while no Comp. Score exceeds 9.00%. The low scores and uneven category results indicate that the fine-grained relation differences evaluated by AirRelComp are not directly captured through existing large-scale pretraining.

We then fine-tune models using only positive AirRelComp pairs under their standard objectives, without controlled negatives. Using the same CLIP-ViP backbone Xue et al. (2023), fine-tuning raises Overall Acc. from 43.21% to 63.45% and Comp. Score from 3.99% to 14.54%, with improvements across all four categories. These gains confirm the benefit of positive-pair fine-tuning on AirRelComp. However, the limited Comp. Score indicates that conventional positive-pair alignment alone does not fully capture the controlled relation differences. RelCompLoss explicitly incorporates the controlled negatives and further raises the fine-tuned CLIP-ViP to 76.37% Overall Acc. and nearly doubles its Comp. Score from 14.54% to 29.23%. Adding SpaSelection further raises these scores to 78.98% and 35.78%, respectively. SpaSelection yields its largest category gain of 9.68 points on Entity-Swap, consistent with its motivation of preserving visual evidence from smaller and spatially dispersed entities. The full framework also outperforms the positive-only fine-tuned

Table 4: Ablation study on learning from controlled negatives on the AirRelComp benchmark.

Variant	Loc-Replace	Dir-Reverse	Stage-Reorder	Entity-Swap	Overall Acc.	Comp. Score
$\mathcal{L}_{\text{InfoNCE}}$	68.79	63.27	55.48	60.22	63.45	14.54
$\mathcal{L}^{\text{Aug}}_{\text{InfoNCE}}$	74.34	65.19	66.98	**62.51**	68.57	20.29
$\mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{Neg}}$	**86.45**	67.85	73.16	60.29	74.79	25.87
$\mathcal{L}_{\text{RelComp}}$	84.04	**71.42**	**79.45**	61.29	**76.37**	**29.23**

baselines in both text-to-video and video-to-text retrieval on the AirRelComp validation set. Full details are provided in Appendix C.3.

5.2.2 Evaluating Large Multi-modal Models with Binary Choice

As shown in Table 3, Qwen3.5-VL-9B achieves the strongest open-source zero-shot performance, with 74.15% Overall Acc. and 33.07% Comp. Score. Notably, our full framework reaches 78.98% Overall Acc. and 35.78% Comp. Score under similarity comparison, exceeding both metrics of zero-shot Qwen3.5-VL-9B under binary choice. Although the task interfaces differ, the results indicate that task-specific learning on AirRelComp enables a much smaller video-text model to achieve competitive performance against a 9B-scale LMM pretrained on large external corpora. The closed-source models achieve higher aggregate zero-shot results. ChatGPT-5.6 sol reaches 86.17% Overall Acc. and 55.77% Comp. Score. However, no closed-source model performs best across all categories. Gemini 3.5 Flash leads Dir-Reverse, ChatGPT-5.6 luna leads Entity-Swap, and ChatGPT-5.6 sol leads Stage-Reorder. This variation further shows that AirRelComp exposes distinct strengths and weaknesses across disruption types rather than the aggregate ranking.

Motivated by the gains from AirRelComp fine-tuning observed in Table 2, we further adapt Qwen3.5-VL-9B, the strongest open-source model, using LoRA. On the held-out test split, adaptation raises its Overall Acc. from 74.15% to 89.82% and its Comp. Score from 33.07% to 62.28%, achieving the best aggregate results among all evaluated LMMs and surpassing the strongest zero-shot closed-source model by 3.65 and 6.51 points, respectively. These gains demonstrate that AirRelComp not only diagnoses fine-grained compositional understanding in LMMs, but also provides effective supervision for adapting them to the relations evaluated by the benchmark.

5.3 Ablation Study

Table 4 compares four objectives for learning from the controlled negatives in AirRelComp. The standard bidirectional $\mathcal{L}_{\text{InfoNCE}}$ uses only positive training pairs. To distinguish the contribution of structured supervision from that of additional negative texts, we introduce a negative-augmentation objective $\mathcal{L}^{\text{Aug}}_{\text{InfoNCE}}$. It adds all controlled negatives to the video-to-text candidate set as generic contrastive negatives, without preserving positive-negative correspondence.

$\mathcal{L}^{\text{Aug}}_{\text{InfoNCE}}$ raises Overall Acc. from 63.45% to 68.57%, confirming that the controlled negatives provide useful supervision. With the same controlled negatives, $\mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{Neg}}$ reaches 74.79%, showing the benefit of preserving positive-negative correspondence. Adding $\mathcal{L}_{\text{Rel}}$ yields 76.37% Overall Acc. and 29.23% Comp. Score, the highest aggregate scores. Relative to $\mathcal{L}_{\text{Neg}}$, it improves Dir-Reverse, Stage-Reorder, and Entity-Swap. These results show that RelCompLoss exploits both positive-negative correspondence and relation identity beyond generic negative exposure.

6 Conclusion

This work addresses fine-grained compositional understanding in continuous multi-entity airport videos. We introduce AirRelComp, which organizes Temporal, Spatial, and Entity Relations under a common event schema and constructs controlled comparisons by changing one designated relation while preserving the remaining event structure. Evaluations across video–text models and LMMs reveal persistent difficulties in distinguishing these relation changes. We further use the controlled comparisons as supervision and develop a framework that improves both relation learning and visual representation. Experiments show clear gains over positive-only fine-tuning and generic negative augmentation, while LoRA adaptation of Qwen3.5-VL-9B confirms that this supervision also benefits LMMs. AirRelComp therefore supports both diagnosing compositional errors and improving the models' performance, providing a basis for faithful video understanding of airport activities.

USE OF LARGE LANGUAGE MODELS

Large language models are used during AirRelComp construction and quality assurance. ChatGPT-5.5 is used to realize predefined relation changes as controlled negative texts. It receives only the human-annotated positive text and the corresponding construction rule, without access to the video; the factual annotation and intended contradiction are therefore determined independently of the model. A separate text-only LLM is used to screen generated pairs for rule compliance, grammatical correctness, and entity–action compatibility. Human reviewers further inspect a subset of the generated pairs and correct residual errors. Large language models were also used for limited language polishing, including grammar and typographical checks, without altering the technical content. Large multi-modal models evaluated or adapted in our experiments are treated as benchmark subjects and follow the fixed evaluation protocols described.

REFERENCES

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.

Alejandro Aparcedo, Akash Kumar, Aaryan Garg, Dalton Pham, Wen-Kai Chen, Anirudh Bharadwaj, Aman Chadha, and Yogesh Rawat. Vista: A benchmark for spatio-temporal interaction understanding in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8411–8422, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13927–13937. IEEE, 2024.

Yang Du, Yuqi Liu, and Qin Jin. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5260–5269, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245.

Florent Fusier, Valery Valentin, François Brémond, Monique Thonnat, Mark Borg, David Thirde, and James Ferryman. Video understanding for complex activity recognition. *Machine Vision and Applications*, 18(3):167–188, 2007.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Koray Kavukcuoglu, Jeff Dean, Oriol Vinyals, and Noam Shazeer. Gemini 3.5: Frontier intelligence with action. `https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/`, May 2026. Google.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, et al. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In *International Conference on Learning Representations*, volume 2024, pp. 40504–40554, 2024.

Dahun Kim, AJ Piergiovanni, Ganesh Mallya, and Anelia Angelova. Videocomp: Advancing fine-grained compositional and temporal alignment in video-text models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29060–29070, 2025.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024.

Yi Li, Kyle Min, Subarna Tripathi, and Nuno Vasconcelos. Svitt: Temporal learning of sparse video-text transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18919–18929. IEEE, 2023.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. Neighborretr: Balancing hub centrality in cross-modal retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9263–9273. IEEE, 2025.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

OpenAI. Introducing gpt-5.4 mini and nano. `https://openai.com/index/introducing-gpt-5-4-mini-and-nano/`, March 2026a.

OpenAI. Gpt-5.5 system card. OpenAI, April 2026b. URL `https://openai.com/index/gpt-5-5-system-card/`.

OpenAI. GPT-5.6 System Card. `https://deploymentsafety.openai.com/gpt-5-6`, July 2026c. Published July 9, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6545–6554. IEEE, 2023.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. Velociti: Benchmarking video-language compositional reasoning with strict entailment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18914–18924. IEEE, 2025.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Phat Thai, Sameer Alam, Nimrod Lilith, and Binh T Nguyen. A computer vision framework using convolutional neural networks for airport-airside surveillance. *Transportation Research Part C: Emerging Technologies*, 137:103590, 2022.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024a.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024b.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14648–14658. IEEE, 2022.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-ViP: Adapting pre-trained image-text model to video-language representation alignment. In *International Conference on Learning Representations*, 2023.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

Yi Yang, Zheng Wang, Xing Xu, Jingkuan Song, and Heng Tao Shen. Gravitation-driven semantic alignment for text video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14946–14956, 2026.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, Yihang Duan, Xinyu Lyu, and Heng Tao Shen. Text-video retrieval with global-localsemantic consistent learning. *IEEE Transactions on Image Processing*, 34:3463–3474, 2025.

Jiahao Zhang, Stephen Gould, and Itzik Ben-Shabat. Vidat—ANU CVML video annotation tool. `https://github.com/anucvml/vidat`, 2020.

Tianxiong Zhang, Zhiqiang Zhang, Xinping Zhu, Boxu Chen, Jiajun Li, and Yuming Zhong. Aircraft engine danger areas incursion detection using keypoint detection and iot. *Alexandria Engineering Journal*, 93:7–21, 2024a.

Xiang Zhang, Yingqi Tang, Maozhang Zhou, Celimuge Wu, and Zhi Liu. Mbsnet: To distinguish motion from stillness for airport traffic safety. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):16817–16830, 2024b.

Zhengyang Zhao, Buhong Wang, Jiwei Tian, Weiwei Jiang, and Yubing Wang. Black-box adversarial attack with nesterov-adax momentum on airport ground vehicle surveillance. *Transactions on Emerging Telecommunications Technologies*, 37(3):e70383, 2026. doi: https://doi.org/10.1002/ett.70383. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.70383`.

Wentao Zhou, Chengtao Cai, Chenming Li, Hao Xu, and Haochen Shi. Ad-yolo: A real-time yolo network with swin transformer and attention mechanism for airport scene detection. *IEEE Transactions on instrumentation and Measurement*, 73:1–12, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, Xiongfei Yao, and Shuaiwei Jiao. Cvbench: Benchmarking cross-video synergies for complex multimodal reasoning, 2026. URL https://arxiv.org/abs/2508.19542.

Weijun Zhuang, Yuqing Huang, Weikang Meng, Xin Li, Ming Liu, Xiaopeng Hong, Yaowei Wang, and Wangmeng Zuo. Cluster-wise spatio-temporal masking for efficient video-language pretraining. *arXiv preprint arXiv:2603.22953*, 2026.

Rolf A Zwaan and Gabriel A Radvansky. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162, 1998.

APPENDIX

Appendix A describes the source videos, event schema and annotation process, controlled negative generation, and quality assurance of AirRelComp. Appendix B provides evaluation and implementation details. Appendix C presents in-depth analyses of AirRelComp and the proposed framework.

A Benchmark Construction Details

This appendix details the construction of AirRelComp, from source videos and human annotation to controlled negative generation and quality assurance. We describe the annotation protocol and event schema, specify the predefined relation changes and generation instructions, and present quality checks and human verification under the same rules, followed by lexical control metrics.

A.1 Source Videos and Event Annotation

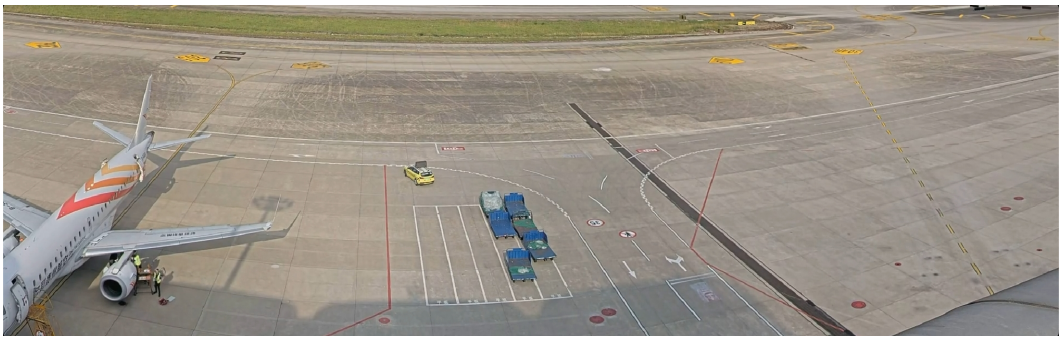
Source videos. AirRelComp is constructed from 118 untrimmed surveillance videos totaling 15.19 hours, recorded from two fixed viewpoints at a civil airport. The videos cover runway, taxiway, and apron activities involving aircraft, ground service vehicles, and personnel. All source videos have a resolution of 2560×806 pixels and a frame rate of 30 fps. Table 5 summarizes the source corpus and annotated clips. Approximately 80% of the annotated clips depict daytime activities, while the remaining clips depict nighttime activities. Figure 3 shows representative examples.

Table 5: Source corpus and annotated clip statistics.

Source corpus	**Value**	**Annotated clips**	**Value**
Source videos	118	Positive pairs	1,544
Total duration	15.19 h	Clips per source, mean	13.09
Frame rate	30 fps	Clips of at most 60 s	93.78%
Resolution	2560 x 806	Duration, mean	23.29 s
Duration, range	2.33-8.37 min	Duration, median	16.80 s
Duration, mean	7.72 min	Duration, range	0.466-237.934 s

Annotators and tool. Eight undergraduate students majoring in engineering participated in annotation. Each annotation was produced by a primary annotator and subsequently reviewed by other annotators, with reviewers adjudicating disagreements. Annotators used Vidat Zhang et al. (2020) in Actions / Video segments mode to select temporal boundaries and enter descriptions. Each annotation records the start time, end time, and natural-language description of the selected span.

The instructions specify participating entities, observable actions, spatial information, and temporal order, transitions, and concurrency, as summarized in Tables 6 and 7. Annotators also received the entity photographs and category references summarized in Table 8. Figure 4 illustrates the scene references and the annotation interface.



(a) Daytime



(b) Nighttime

Figure 3: Representative daytime and nighttime scenes in AirRelComp.

Annotation content. The guidelines define the entities, operating areas, spatial expressions, actions, and temporal information to be recorded. Table 6 summarizes this annotation vocabulary and scope; visually observable cases outside the reference vocabulary are resolved during review.

Table 6: Content covered by the annotation guidelines.

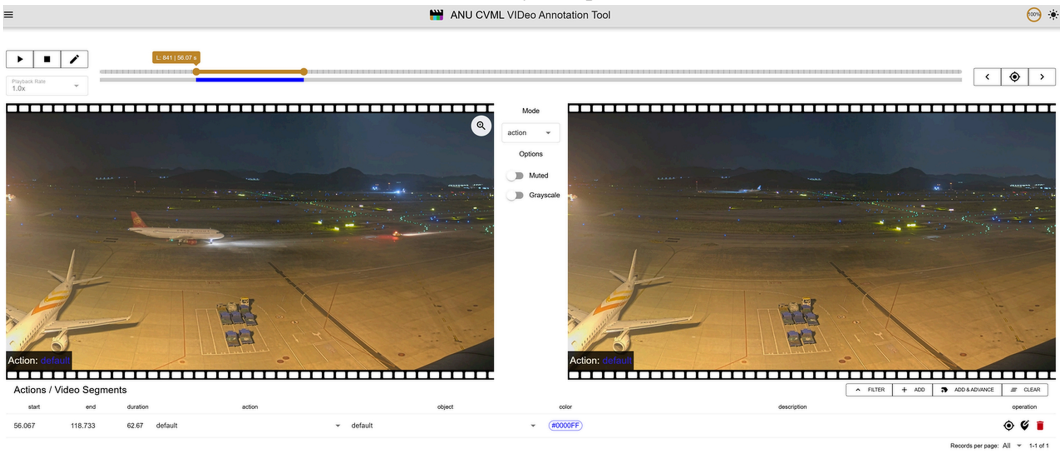
Dimension	**Categories, vocabulary, and instructions**
Participating entities	Aircraft; ground service vehicles; personnel.
Operating areas	Runway; taxiway; apron.
Motion directions	From left to right; from right to left; from the lower to the upper part of the image; from the upper to the lower part of the image.
Locations and references	Left, right, middle, upper, and lower monitored area; near, to the left of, or to the right of a clearly identifiable entity or operating area. Avoid very small objects as spatial references.
Personnel actions	Inspecting or maintaining aircraft; guiding aircraft or vehicles; carrying baggage; assisting refueling, catering, cleaning, or baggage services; moving scaffolding.
Aircraft actions	Taxiing; stopping; taking off; landing; turning between a runway and taxiway; moving from a taxiway to the apron. Record maintenance received when visible.
Vehicle actions	Guiding aircraft; supplying fuel or food; towing; transporting baggage; moving; stopping or waiting after movement; providing visible ground services.
Entity participation	Specify entity type and count, the activity, its spatial context, and the entity or object involved when applicable.
Temporal development	Describe stage transitions, simultaneous activities, changes in participation, and the final state or exit from the monitored area.
Optional modifiers	Use slow or fast when supported by visible motion. Prefer a specific visible action to a vague description such as providing services.



(a) Apron viewpoint.



(b) Runway viewpoint.



(c) Vidat annotation.

Figure 4: Scene references and the interface used for temporal annotation. The two views identify runway, taxiway, and apron regions. The Vidat panel shows the selected temporal interval and its associated description.

Table 7: Operational instructions for producing activity descriptions.

Step or case	**Instruction**
Inspect the video	Watch the source video before annotating to understand the activity sequence and distinguish the participants.
Set temporal boundaries	Place the start boundary at the onset of the described activity and the end boundary at its completion.

Step or case	Instruction
Use spatial references	Describe location and movement relative to the camera view. Combine a scene or location phrase with a motion direction when appropriate.
Describe concurrent activity	Represent the intervals before, during, and after an overlap when the set of active participants changes. Preserve which entity performs each action.
Describe scene transitions	Name both the departure and destination areas, followed by the subsequent motion when visible.
Distinguish similar entities	Use spatial qualifiers first; add entity counts or a longer description when location alone is insufficient.
Include the ending	Record the final state or exit boundary. Describe stopping or waiting as an activity stage following visible movement.
Resolve unclear cases	Discuss activities or entity categories that cannot be described confidently. Reviewers adjudicate disagreements before finalization.

Entity identification. The visual reference distinguishes aircraft and personnel from ground vehicles and service equipment. Table 8 lists the entity terms and summarizes visual cues from the supplied examples. Figure 5 presents the corresponding visual references used to support entity identification in descriptions.

Table 8: Entity terms and visual cues summarized from the identification reference supplied to annotators.

Entity term	Visual cue in the reference examples
Aircraft	Fuselage, wings, tail, and landing gear.
Ground crew member	People visible near aircraft or service vehicles.
Catering truck	Truck with a tall enclosed service body, also shown in a raised position.
Baggage cart	Low trailer shown separately or coupled behind a tug.
Baggage tractor	Compact towing vehicle, shown with attached baggage carts.
Aircraft tug	Broad, low towing vehicle shown close to an aircraft.
Follow-me vehicle	Small vehicle with conspicuous body markings and roof equipment.
Shuttle bus	Long passenger vehicle with an extended row of windows.
Baggage belt loader	Vehicle with a long inclined conveyor.
Fueling truck	Service vehicle with a large tank and visible service equipment.
Potable water truck	Compact service truck with a tank behind the cab.
Passenger stair truck	Mobile inclined stairway with an upper access platform.

Representative descriptions. Table 9 illustrates how the guidelines encode location, movement, transitions, concurrency, participation changes, and final states.

Table 9: Examples of activity descriptions organized from the annotation guidelines.

Focus	Example description
Location and movement	On the runway, an aircraft slowly taxis from right to left.
Transition between areas	An aircraft moves from the runway to the taxiway, then taxis toward the lower part of the image.
Concurrent activity	On the taxiway, a follow-me vehicle drives from left to right while an aircraft follows it in the same direction.
Participation changes	On the left side of the apron, two personnel finish maintenance while three continue working on a parked aircraft.
Final state and exit	A tractor enters from the right side of the apron and leaves through the lower boundary of the central area.

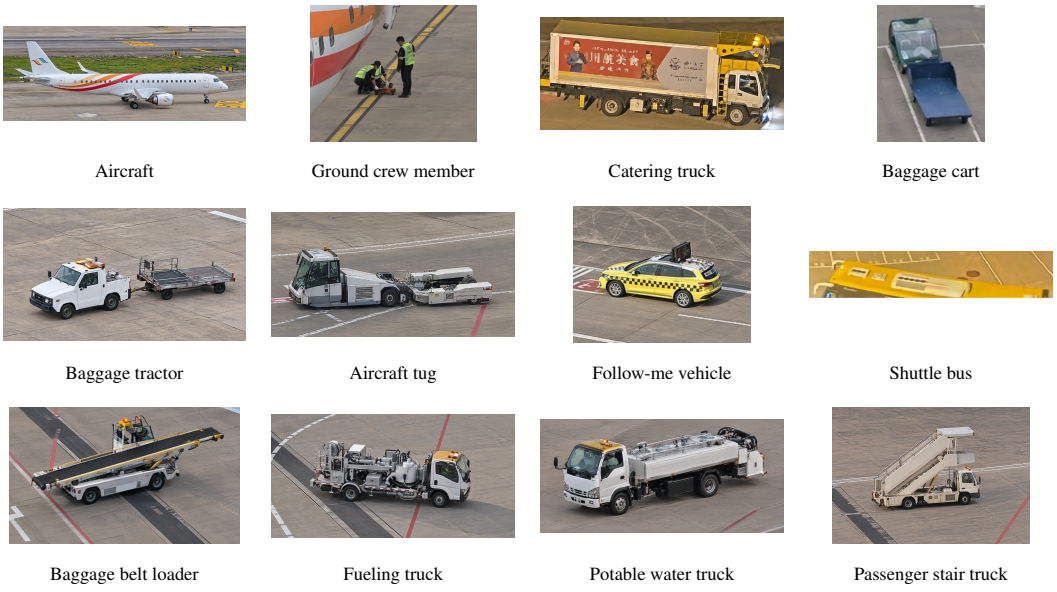


Figure 5: Visual entity references supplied for annotation, organized by the entity terms in Table 8.

Event schema. Following cognitive theory that represents events along multiple situational dimensions, including time, space, and participating entities Zwaan & Radvansky (1998), annotators track aircraft, vehicles, and personnel across space and time, recording visually observable entities and actions together with their locations, motion directions, temporal stage transitions, and participation in sequential and concurrent events. For each annotated activity span, annotators mark its start and end timestamps and write a positive text t_i^+ for the corresponding video segment v_i. The resulting positive text expresses a relational event structure, denoted by $\mathcal{S}_i^+$:

$$\mathcal{S}_i^+ = (\mathcal{C}_i; \mathcal{R}_{T,i}, \mathcal{R}_{S,i}, \mathcal{R}_{E,i}) \,, \tag{8}$$

where $\mathcal{C}_i$ denotes the event content in t_i^+. $\mathcal{R}_{T,i}$ encodes activity ordering, transitions, and concurrency, $\mathcal{R}_{S,i}$ encodes entity locations and movement, and $\mathcal{R}_{E,i}$ encodes participation assignments between entities and event units. Human annotation and LLM-assisted standardization yield temporally aligned positive video–text pairs (v_i, t_i^+) that preserve the observed relational event structure.

A.2 Controlled Negative Generation

Rule design. Before generation, we define four controlled relation changes: Spatial Location Replacement (Loc-Replace), Spatial Direction Reversal (Dir-Reverse), Temporal Stage Reordering (Stage-Reorder), and Entity Participation Swapping (Entity-Swap). Each rule specifies its target relation, applicability conditions, permitted transformation, and preservation constraints. If no eligible relation is present, no negative of that type is generated. Table 10 summarizes the targeted changes and preserved content.

Table 10: Targeted changes and preserved content for the four controlled compositional disruptions.

Disruption	**Targeted change**	**Content to preserve**
Loc-Replace	Replace one reported location of an entity or activity.	Entity identities, activities, temporal organization, and other spatial content.
Dir-Reverse	Reverse motion direction, exchange path endpoints, or replace a motion target as permitted.	Entity identities, activities, temporal organization, and other spatial content.

Disruption	Targeted change	Content to preserve
Stage-Reorder	Change the order of complete activity stages performed by the same entity.	The activity content, participating entities, and spatial information within each stage.
Entity-Swap	Change participation assignments between existing entities and activities.	The existing activities and their temporal and spatial organization.

Construction resources. The following resources specify how each predefined rule is applied. They define the permitted substitutions, applicability conditions, and exclusions while requiring all non-target event content to remain unchanged. A valid transformation may require coordinated edits to multiple text spans, such as exchanging both endpoints of a directed path.

Loc-Replace. This rule changes one reported location of an entity or activity, including its appearance, entry, exit, or stopping position. Eligible locations also include positions relative to an aircraft. Table 11 defines the replacement vocabulary. Selection prioritizes entity local part or relative side attributes, current or final positions, appearance or entry or exit positions, positions relative to the surrounding area, and finally broad scene regions. A broad region is replaced only when no finer eligible location is available. The rule preserves entity identities, activities, temporal order, motion paths, and other spatial content. It is inapplicable when the candidate phrase denotes only a motion source, target, or direction, or when no meaningful replacement is available.

Table 11: Loc-Replace vocabulary and application constraints. Bidirectional arrows permit alternatives in both directions; single arrows permit only the indicated substitution.

Location expression	Permitted replacement	Application constraint
Scene region	runway ↔ taxiway; runway ↔ apron; taxiway ↔ apron	Use only for a broad scene region associated with an entity or activity when no finer location is available.
Relative position	above ↔ below	Preserve the reference region and the rest of the spatial relation.
Horizontal side	left side ↔ right side	Change only the selected location.
Vertical side	upper side ↔ lower side	Retain the reference region.
Diagonal side	upper-left side ↔ lower-right side; upper-right side ↔ lower-left side	Apply within a location or state description.
Middle side	middle-left side ↔ middle-right side	Preserve the rest of the location description.
Central position	middle → left side or right side	Adjust the preposition only as required for natural wording.
Monitored-area position	at the left side of the monitored area ↔ at the right side of the monitored area	Current, appearance, or stopped position.
Appearance, entry, or exit side	from the left side of the monitored area ↔ from the right side of the monitored area	Applies to an appearance, entry, leaving, or exit side, not an ordinary movement source.
Entry side	enters from the left side ↔ enters from the right side	Preserve any subsequent motion.
Stopped position	stops at the left side ↔ stops at the right side	Preserve the stopping action.
Corner position	at the lower-right side ↔ at the upper-left side; at the lower-left side ↔ at the upper-right side	Keep the same reference region.
Aircraft part	aircraft's nose ↔ aircraft's tail; aircraft nose ↔ aircraft tail; aircraft nose area ↔ aircraft tail area	Use only for a location or state description; motion targets use Dir-Reverse.
Local relation	near the aircraft's nose ↔ near the aircraft's tail; at the aircraft nose ↔ at the aircraft tail	Retain *near* or *at*; do not introduce *away from*.

Location expression	Permitted replacement	Application constraint
Aircraft side	left side of the aircraft ↔ right side of the aircraft	Preserve the aircraft identity and activity.
Stationary-aircraft side	at the left side of a stationary aircraft ↔ at the right side of a stationary aircraft	Preserve *stationary* and the participating entity.
Aircraft wing	aircraft's left wing ↔ aircraft's right wing; left wing of the aircraft ↔ right wing of the aircraft	Determine whether the phrase describes a location or a motion target, not merely whether it contains *wing*.
Parked-aircraft position	parked aircraft on the right side ↔ parked aircraft on the left side	Preserve the parked state and existing activity.
Parked-aircraft center	parked aircraft on the right side → parked aircraft in the middle; parked aircraft on the left side → parked aircraft in the middle	Preserve the reference area and aircraft identity.
Central-area form	In the middle of REGION → On the left side of REGION; In the middle area of REGION → On the left side of REGION	Replace the complete location expression and retain REGION.
Side-area form	At the right-side area → At the left-side area	A location at which an activity occurs.
Partial diagonal edit	lower-right area/part → lower-left area/part	Change only the horizontal component of the location.
Boundary alternative	lower-middle area/side → upper-right area/side	Use only as a permitted boundary alternative, not as a lexical antonym.
Other entry and exit expressions	walk in / leave / exit from SIDE → the same event with an allowed alternative SIDE	Use the applicable side mapping above; retain other stages.

The reference object or region remains unchanged except for the explicitly permitted region substitutions. Natural central expressions include *in the middle of the apron*, *in the middle of the monitored area*, and *in/at the middle part of the aircraft*. Necessary preposition changes do not license a change from *near* to *away from*. A motion verb elsewhere in the description does not disqualify a separate location expression.

Dir-Reverse. This rule changes a movement direction, turn, motion target, or path source and destination using the predefined mappings in Table 12. Eligible movements include walking, driving, taxiing, flying, approaching, and moving an object toward a target. Only the selected component of the movement changes; exchanging path endpoints changes both the source and destination together. The rule preserves the moving entity, activity, temporal order, and unrelated locations. It is inapplicable to static positions, appearance, entry, or exit sides, objects of maintenance without a movement relation, and symmetric connections such as *connects A with B*.

Table 12: Dir-Reverse mappings and application constraints. Bidirectional arrows permit alternatives in both directions; single arrows permit only the indicated substitution.

Operation	Replacement	Constraint
Horizontal reversal	left ↔ right; from left to right ↔ from right to left	Edit a movement component, not every directional word in the sentence.
Vertical reversal	upward ↔ downward; toward lower side ↔ toward upper side	Preserve the movement predicate.
Turning reversal	turns left ↔ turns right	Do not replace turning with another action.
Compound direction	toward front-right side ↔ toward rear-left side	Preserve the remaining movement description.
Diagonal target	toward upper-left side ↔ toward lower-right side; toward upper-right side ↔ toward lower-left side	Change the motion target, not an appearance, entry, or exit side.

Operation	Replacement	Constraint
Central target	toward middle side → toward left side or toward right side	A target substitution, not a literal direction reversal.
Aircraft target	aircraft nose ↔ aircraft tail, within *toward*, *to*, or *approach*	Keep the reference aircraft and the original movement relation; do not introduce *away from*.
Wing target	left wing ↔ right wing	Applies to a motion target; static positions use Loc-Replace.
Reference objects as endpoints	from OBJECT A toward OBJECT B → from OBJECT B toward OBJECT A	Both objects must already occupy the endpoints of the same path.
Region endpoints	from REGION A toward REGION B → from REGION B toward REGION A	Retain the moving entity and movement predicate.
General endpoints	moves from X to Y → moves from Y to X	Exchange the two endpoints together and retain their modifiers.

Loc-Replace applies when the edited phrase describes where an entity appears, enters, leaves, exits, stops, or performs an activity. Dir-Reverse applies when it describes movement direction, a turn, a motion target, or path endpoints. Other movement stages in the sentence do not change this distinction. Table 13 provides examples.

Table 13: Loc-Replace changes a location; Dir-Reverse changes movement direction, a motion target, or path endpoints. Only the bold component is replaced.

Edited expression	Disruption type
On the **right** side of the apron → On the **left** side of the apron	Loc-Replace
crew at the aircraft **nose** moves → crew at the aircraft **tail** moves	Loc-Replace
crew walks toward the aircraft **nose** → crew walks toward the aircraft **tail**	Dir-Reverse
crew from the aircraft's **left wing** moves → crew from the aircraft's **right wing** moves	Dir-Reverse
appears from **right**, then flies toward upper-left → appears from **left**, then flies toward upper-left	Loc-Replace
appears from right, then flies toward **upper-left** → appears from right, then flies toward **lower-right**	Dir-Reverse
stops near the aircraft **tail** → stops near the aircraft **nose**	Loc-Replace

Stage-Reorder. This rule reorders two or more complete activity stages performed by the same entity. Temporal connectors or unambiguous ordered clauses must establish their order. A local permutation changes that order while preserving each stage's activity, participating entities, and spatial content. Table 14 specifies the permitted operations. The rule is inapplicable when the entity's identity across stages or the stage order is uncertain, activities form an unordered list, or the permutation is semantically equivalent.

Table 14: Stage-Reorder operations and applicability constraints.

Rule component	Permitted operation	Constraint
Eligibility	Reorder at least two clearly ordered activity stages performed by the same entity.	The entity, stage boundaries, and temporal order must be unambiguous.
Temporal cues	Identify order from *then*, *next*, *subsequently*, *afterward*, *after that*, *eventually*, *finally*, *before*, *after*, or an unambiguous ordered comma sequence.	A connector alone is insufficient, and unordered activity lists are excluded.

Rule component	Permitted operation	Constraint
Complete activity stage	Move each stage together with its action, adverbs, objects, quantities, direction, source, and target.	Do not split a stage or alter its spatial content or participation assignments.
Two stage permutation	A then $B \rightarrow B$ then A; A before $B \rightarrow B$ before A; A after $B \rightarrow B$ after A.	Interpret the original temporal relation before applying the permutation.
Multiple stage permutation	Apply one clear local permutation, such as $A \rightarrow B \rightarrow C \;\Rightarrow\; A \rightarrow C \rightarrow B$.	Preserve every stage exactly once and leave unrelated event sequences unchanged.
Connector and grammar repair	Retain existing connectors where possible; insert a minimal *then* or adjust finite and gerund forms when required.	Repairs must not introduce another semantic change.
Relation boundary	Reorder only sequential stages performed by the same entity.	Do not reorder stages performed by different entities or convert *meanwhile* or *at the same time* into a sequential relation.
Event preservation	An existing exit stage may precede another existing stage after reordering.	Do not add an imagined reentry or any other event to improve plausibility.
Exclusion	Produce no negative when the entity's identity across stages or their order is uncertain, the permutation is equivalent, or event content would be lost.	Stage-Reorder may be inapplicable independently of the other three types.

Entity-Swap. This rule changes participation assignments between existing entities and activities. It requires at least two distinguishable entities or groups of entities and two existing activities whose participation assignments can be changed. Each activity is considered with its participating entities, action or state, location, movement, targets or objects, and stated quantities. Entities may be distinguished by category, explicit quantity, role, a location associated with the entity, or a stable identifying description. The rule reassigns entities to existing activities while preserving each activity's spatial content and the overall temporal order, with only the specified entity–action compatibility edits. It is inapplicable when the entities cannot be distinguished, the assignment is equivalent, or a valid description would require inventing an entity or activity. Table 15 lists expressions for identifying entities and spatial references already present in the positive text.

Table 15: Entity and spatial reference expressions used for Entity-Swap. Only entities already present in the positive text are eligible.

Term category	Terms
Aircraft	aircraft
Personnel	ground crew member; staff member; worker; person
Vehicles and equipment	follow-me vehicle; baggage cart; baggage tractor; baggage tug; baggage belt loader; refueling truck; catering truck; aircraft tug; pushback tug; ground support vehicle; de-icing truck; sweeping vehicle; patrol vehicle; fire truck; maintenance vehicle; potable water truck; shuttle bus; passenger stair truck
Aircraft parts	aircraft nose; aircraft tail; aircraft left wing; aircraft right wing
Reference and quantity expressions	one; another; two; three; a total of . . .

Table 16 specifies the permitted participation assignments, compatibility constraints, and exclusions. Separate clauses, a shared location, or different objects of maintenance alone do not distinguish otherwise identical entities. Merely exchanging *one* and *another*, reversing a symmetric connection, or swapping synonymous action descriptions does not change participation assignments.

Table 16: Entity-Swap participation assignments and compatibility constraints.

Constraint	Permitted operation or restriction
Participation assignment	Reassign distinct existing entities to existing activities. Prefer a clear exchange between two entities; use a larger reassignment only when necessary to change participation assignments.
Same-category entities	Permit exchanges when the descriptions distinguishing the entities remain interpretable and the resulting participation assignment actually differs. Do not merely relabel indistinguishable entities.
Object and participation roles	An existing target or object may exchange roles with a participating entity in another activity. Retain all existing participating entities, including any personnel performing maintenance.
Existing quantities	Change participation assignments of existing groups without inventing a quantity. Repeated or nested group mentions do not justify adding their counts to infer a total.
Motion compatibility	For ground motion, use *taxi/move* for aircraft, *drive/move* for vehicles, and *walk/move* for personnel when adaptation is necessary. Retain compatible actions such as pulling, pushing, or towing. Infrastructure and aircraft parts must not acquire human actions.
State compatibility	Remove a conflicting *parked* modifier only when needed for the reassigned entity to participate in an activity involving movement. Do not otherwise remove state or identity information.
Complete entity names	Preserve compound names such as *aircraft tug*; do not shorten them or infer an entity category from an ambiguous name.
Preserved structure	Keep temporal order, temporal connectors, and the spatial content of each activity. Do not add participants, groups, activities, or an independent change in direction or region.
Invalid assignments	Exclude equivalent target swaps between indistinguishable entities, symmetric connections, and synonymous action descriptions. Exchanging the entity performing a single activity with an existing target or object in that activity does not satisfy the requirement for two existing activities.

Constrained generation. We use ChatGPT-5.5 OpenAI (2026b) to apply the designated rule to each positive text without video input. The positive annotation supplies the factual description, and the rule determines the permitted contradiction. The model identifies an eligible relation under that rule and realizes the specified change in language. It preserves unrelated wording and makes only necessary grammatical or explicitly permitted compatibility edits.

Each positive text is considered independently for all four types, with at most one accepted negative per type. Each negative remains associated with the original positive annotation and its temporal span. If the required relation is absent, the change is semantically equivalent, or valid wording requires guessing event content or extensive rewriting, no negative of that type is generated. Inapplicability of one type does not prevent generation of another.

Generation instructions. We summarize the instructions for each type below in a compact form. The referenced rule tables supply the permitted transformations.

Loc-Replace instruction

Input. Positive text; designated type Loc-Replace; the rule in Table 11.

Task. Select one eligible location expression and apply one permitted replacement.

Constraints. Preserve entity identities, quantities, activities, temporal order, motion paths, and other locations. Retain the reference object and relation except for the permitted region substitutions; make only minimal grammatical edits.

Eligibility. Generate no negative if no clear location expression exists, the change affects a motion path or another relation, or the replacement is semantically equivalent.

Output. One controlled negative text when applicable.

Dir-Reverse instruction

Input. Positive text; designated type Dir-Reverse; the rule in Table 12.

Task. Apply one permitted direction reversal, motion target substitution, or exchange of path endpoints.

Constraints. Preserve the moving entity, activity, quantities, temporal order, and unrelated locations. Exchange both endpoints together when required. Preserve the reference aircraft for a local motion target and make only minimal grammatical edits.

Eligibility. Generate no negative without a directional motion relation, for a symmetric connection, or when the edit changes only a static location or an appearance, entry, or exit position. Exclude equivalent changes and edits requiring a new entity or activity.

Output. One controlled negative text when applicable.

Stage-Reorder instruction

Input. Positive text; designated type Stage-Reorder; the rule in Table 14.

Task. Identify at least two ordered activity stages performed by the same entity and apply one local permutation that changes their temporal order.

Constraints. Retain every complete stage exactly once, including its action, participating entities, quantities, direction, source, and target. Preserve unrelated sequences and concurrency. Retain connectors where possible; make only necessary grammatical changes or insert a minimal *then*. Do not invent an event to make the reordered sequence plausible.

Eligibility. Generate no negative when stage boundaries, the entity's identity across stages, or their order are unclear, when activities form an unordered list, or when the permutation is equivalent.

Output. One controlled negative text when applicable.

Entity-Swap instruction

Input. Positive text; designated type Entity-Swap; the rule in Table 16.

Constraints. Preserve entity identities, complete names, existing quantities, activities, temporal order, and spatial content. Apply only the specified motion or state compatibility edits and minimal grammatical adjustments. Do not introduce an entity, activity, or additional relation change.

Eligibility. Generate no negative when entities are indistinguishable, participation assignments are equivalent, or the edit merely swaps a symmetric connection or synonymous action descriptions.**Task. Identify at least two distinguishable entities or groups of entities and two existing activities, then change their participation assignments.**

Constraints. Preserve entity identities, complete names, existing quantities, activities, temporal order, and spatial content. Apply only the specified motion or state compatibility edits and minimal grammatical adjustments. Do not introduce an entity, activity, or additional relation change.

Eligibility. Generate no negative when entities are indistinguishable, participation assignments are equivalent, or the edit merely swaps a symmetric connection or synonymous action descriptions.

Output. One controlled negative text when applicable.

A.3 Quality Assurance and Lexical Control

Quality checks. We use a separate text-only LLM checking task to compare each generated negative with its positive annotation under the assigned construction rule. The checker receives the positive text, negative text, assigned disruption type, and corresponding predefined rule, without video or images. It locates the edited spans, distinguishes the intended relation change from necessary grammatical or compatibility edits, and checks whether the candidate satisfies the criteria in Table 17.

Table 17: Criteria used to check generated negatives against their positive annotations and predefined construction rules.

Aspect	Check
Semantic difference	The negative contradicts its positive annotation in the designated relation, rather than merely repeating it, correcting grammar, paraphrasing it, or swapping a symmetric relation.
Rule compliance	The edited spans implement the assigned Loc-Replace, Dir-Reverse, Stage-Reorder, or Entity-Swap rule, while unrelated event content and the other relation dimensions remain unchanged.
Entities and quantities	Participants are neither invented nor removed, and compound entity names remain intact. Entity-Swap may change participation assignments of existing groups but may not introduce new quantities.
Activities and compatibility	Activities are preserved except for the minimal compatibility edits permitted by Entity-Swap. Infrastructure or aircraft parts must not acquire incompatible human actions.
Temporal and spatial structure	Stage-Reorder changes the order of complete activity stages performed by the same entity while preserving each stage's activity, participating entities, and spatial content. The other rules preserve stage order. For Spatial Relations, the checker determines whether the edited phrase describes a location or movement.
Entity distinction	Entity-Swap must change participation assignments between distinguishable existing entities or groups; exchanging targets between activities performed by otherwise indistinguishable entities is not accepted.
Language quality	The negative is checked for damaged phrases, repeated words, missing components, agreement errors, corrupted entity names, and other grammatical defects.

For negatives targeting Spatial Relations, the checker follows Table 13: appearance, entry, exit, and stopping positions are checked under Loc-Replace, whereas movement directions, turns, motion targets, and path endpoints are checked under Dir-Reverse. Candidates with an invalid or ambiguous relation change, incorrect disruption type, unintended additional change, language problem, or entity–action incompatibility are flagged for human review. The text-only checker does not rewrite candidates, and its decision is not treated as video-based verification.

Checking instruction. The following compact instruction applies the criteria independently of generation:

> **Input.** Positive text; generated negative; assigned type from Loc-Replace, Dir-Reverse, Stage-Reorder, or Entity-Swap; the corresponding construction rule.
>
> **Task.** Identify all edited spans and determine whether they implement exactly the assigned relation change.
>
> **Criteria.** Apply Table 17. Check the permitted transformation, preservation of other content, a meaningful semantic difference, grammatical correctness, and entity–action compatibility. Treat necessary grammatical or permitted compatibility edits separately from additional relation changes.
>
> **Output.** Flag noncompliant or uncertain comparisons for human review, identifying the affected text and reason. This checking step does not rewrite the negative.

Human review and correction. Human reviewers inspect 25% of generated comparisons under the same criteria, combining spot checks with targeted inspection of flagged cases. When a recurring violation of a predefined rule is identified, other comparisons exhibiting the same pattern are checked and corrected when needed. Corrections are restricted to the affected content and preserve all unrelated event information.

For any required rewrite, correction starts from the original positive text and re-applies only the designated rule, preventing multiple disruptions from accumulating in one negative. Table 18 summarizes the treatment of reviewed cases.

Table 18: Treatment of comparisons after human review under the predefined construction rules.

Reviewed case	Treatment
Valid comparison	Retain the comparison without rewriting it for stylistic consistency.
Valid negative with an incorrect disruption label	Correct the label only when the implemented relation change is unambiguous; otherwise refer the comparison for further review.
Minor language or entity-name error	Apply only the minimal repair needed to preserve the intended controlled negative. Do not broadly paraphrase or guess an entity.
Invalid or equivalent comparison without a safe correction	Remove the comparison when it does not realize a valid controlled relation change or cannot be corrected without introducing unsupported content.
Uncertain meaning or type	Keep the comparison under review rather than assigning an unsupported pass or correction.

Test set verification. To strengthen the reliability of benchmark evaluation, four reviewers manually verify all 766 comparisons across 309 video clips in the test split. For each comparison, the reviewers confirm that the positive text accurately describes the video and that the negative contradicts it through the designated relation change. They also check preservation of the remaining event content, grammatical correctness, and entity–action compatibility. Verification outcomes are recorded for each comparison. Of the 766 comparisons, 762 (99.48%) pass without revision, while the remaining four are corrected and pass rechecking against the video. No comparisons are excluded.

Lexical control. We assess lexical control using Token Jaccard and positive vocabulary retention. Let V^+ and V^- denote the token sets of the positive and negative texts. For each comparison,

$$J = \frac{|V^+ \cap V^-|}{|V^+ \cup V^-|}, \qquad R = \frac{|V^+ \cap V^-|}{|V^+|}, \tag{9}$$

where J measures token-set overlap and R measures the proportion of unique positive-text tokens retained in the negative. The values reported in Table 1 are averaged over all controlled comparisons.

B Evaluation and Implementation Details

This appendix provides the evaluation and implementation details needed to reproduce our experiments. It covers the evaluation metrics, training and evaluation configurations, baseline initialization, and the deterministic implementation of SpaSelection.

B.1 Evaluation Metrics

We evaluate each available triplet $(v_i, t_i^+, t_{i,k}^-)$ as one controlled comparison. Let $\mathcal{I}_c$ contain the comparison indices (i, k) for disruption type c in the evaluated split, with $N_c = |\mathcal{I}_c|$. The indices $c = 1, 2, 3, 4$ correspond to Loc-Replace, Dir-Reverse, Stage-Reorder, and Entity-Swap, respectively. We assign each comparison a correctness indicator:

$$z_{i,k} = \begin{cases} \mathbb{I}\left[s(v_i, t_i^+) > s(v_i, t_{i,k}^-)\right], & \text{similarity comparison,} \\ \mathbb{I}\left[\widehat{o}_{i,k} = o_{i,k}^+\right], & \text{binary choice,} \end{cases} \tag{10}$$

where $\mathbb{I}[\cdot]$ equals one when its condition holds and zero otherwise. Here, $s(v, t)$ denotes the video–text similarity, $\widehat{o}_{i,k}$ is the LMM response, and $o_{i,k}^+ \in \{\texttt{A}, \texttt{B}\}$ is the option assigned to the positive text. We count similarity ties and LMM responses of `tie` as incorrect.

Overall Accuracy. For each disruption type, we compute the category accuracy as

$$a_c = \frac{1}{N_c} \sum_{(i,k) \in \mathcal{I}_c} z_{i,k}, \qquad c \in \{1, 2, 3, 4\}. \tag{11}$$

We report category accuracies as $100a_c\%$. Overall Accuracy (Overall Acc.) measures the proportion of correct predictions across all controlled comparisons:

$$\text{Overall Acc.} = 100\,\frac{\sum_{c=1}^{4} N_c a_c}{\sum_{c=1}^{4} N_c}. \tag{12}$$

This aggregation gives each comparison equal weight and weights each category by its number of comparisons.

Comprehensive Score. Following VideoComp Kim et al. (2025), we compute the Comprehensive Score (Comp. Score) as the product of the four category accuracies:

$$\text{Comp. Score} = 100 \prod_{c=1}^{4} a_c. \tag{13}$$

We use the unrounded accuracies $a_c \in [0, 1]$ and report both aggregate metrics as percentages. The product favors consistently high accuracy across all four disruption types. Each category uses its own set of available comparisons; the calculation does not require every video to have negatives of all four types.

B.2 Training and Evaluation Configuration

Text relation queries. The text relation module contains one shared cross attention layer with hidden dimension 512 and eight attention heads. Three learned query vectors attend to the final text token states as keys and values, with padding positions masked. We add the attention output to the queries and apply layer normalization. A feedforward network with dimensions $512 \to 512 \to 512$ and GELU activation follows, with another residual connection and layer normalization. A shared linear projection without bias maps each output to the embedding space of dimension 512, followed by ℓ_2 normalization. We initialize the query vectors from $\mathcal{N}(0, 0.02^2)$.

Video–text framework. We initialize our framework from the released CLIP-ViP-B/32 checkpoint and sample 64 frames per video at 224×224 resolution with a maximum text length of 76. SpaSelection is applied after the fourth visual Transformer block, retaining $K = 24$ patch tokens per frame with $C = 4$ clusters. The three patch-scoring terms are weighted by 1.0, 0.5, and 0.5. For RelCompLoss, we set $\rho = 0.3$, $\alpha = 0.5$, $\beta = 0.2$, and the relation temperature to 0.07. We train for 10 epochs with AdamW using a batch size of 8, learning rates of 10^{-6} for the backbone and 10^{-5} for newly introduced parameters, weight decay of 0.2, cosine decay, and a warm-up ratio of 0.01. Video–text baselines follow their released input configurations.

LMM evaluation and adaptation. For LMM evaluation, all models receive the same cached sequence of 64 uniformly sampled frames, with the longer image side limited to 448 pixels, and use the same prompt, A/B assignment protocol, and deterministic decoding. We adapt Qwen3.5-VL-9B with LoRA while freezing its vision encoder. LoRA is applied to the language-model linear layers and the multi-modal aligner with rank 16, scaling factor 32, and dropout 0.05. Training runs for three epochs with an effective batch size of 16, learning rates of 10^{-4} and 2×10^{-5} for the language model and multi-modal aligner, respectively, AdamW with weight decay 0.01, cosine scheduling, and a warm-up ratio of 0.05.

B.3 Baseline Models and Initialization

For similarity comparison, we evaluate video–text models under zero-shot and fine-tuning settings. In zero-shot evaluation, InternVideo2 is pretrained on a mixture including WebVid and InternVid Wang et al. (2024b) and uses 4 frames per video. ViFi-CLIP is pretrained on Kinetics-400 Rasheed et al. (2023) and uses 32 frames. ViCLIP and LanguageBind-Video are pretrained on InternVid-10M-FLT Wang et al. (2024a) and VIDAL-10M Lin et al. (2024), respectively, and both use 8 frames. CLIP-ViP is pretrained on HD-VILA-100M Xue et al. (2023) and uses 64 frames. For fine-tuning, CLIP4Clip Luo et al. (2022), NeighborRetr Lin et al. (2025), and GLSCL Yang et al. (2026) are initialized from CLIP ViT-B/32 Radford et al. (2021). CLIP4Clip uses up to 52 frames per video, while NeighborRetr and GLSCL use 12 and 64 frames, respectively. TV-ProxyNet Zhang et al. (2025) is initialized from CLIP-ViP-B/32 Xue et al. (2023) and uses 12 frames. The fine-tuned

CLIP-ViP baseline, CLIP-ViP with RelCompLoss, and the full framework all use 64 frames per video during training and evaluation.

For binary choice, we evaluate open-source LMMs including Qwen2.5-VL-7B Hui et al. (2024), Qwen3-VL-8B Bai et al. (2025), Qwen3.5-VL-4B/9B Team (2026), InternVL3-8B Zhu et al. (2025), VideoChat-R1.5 Yan et al. (2026), and Owl-Con Bansal et al. (2024), as well as the closed-source Gemini and ChatGPT models. All LMMs are evaluated under the unified protocol.

B.4 SpaSelection Implementation

Score normalization. The neighborhood $\mathcal{N}_r(n)$ contains positions within a radius of $r = 1$ in the previous frame. The change and displacement scores are set to zero for the first frame. Each score is independently normalized over the P patches in every frame using min-max normalization $\mathcal{N}_t(\cdot)$. If its range does not exceed 10^{-6}, all normalized values are set to zero.

Deterministic clustering and selection. We cluster the normalized patch features independently within each frame using deterministic cosine K-means before completing the selection by score. The highest scoring patch initializes the first center, and each subsequent center is chosen as the patch farthest from its nearest existing center. Patches are assigned to the center with the highest cosine similarity. Non-empty centers are updated using the normalized mean of their members, while empty centers remain unchanged. We perform five iterations of assignment and center update without early stopping, followed by a final assignment using the updated centers. After clustering, we retain the highest scoring patch from each non-empty cluster and fill the remaining positions in descending order of $s_{t,n}$ until K patches are selected. Ties are resolved by the lower patch index, and the selected indices are sorted to recover the original patch order.

C In-depth Analysis

We provide further analyses that complement the main evaluation of AirRelComp. Appendix C.1 examines textual artifacts and language bias in the controlled comparisons. Appendix C.2 analyzes the sensitivity of SpaSelection to the number of selected patches and clusters. Appendix C.3 evaluates whether learning from controlled relation changes remains compatible with standard video–text retrieval. Appendix C.4 provides a qualitative analysis of model behavior under the four controlled relation changes.

C.1 Textual Artifacts and Language Bias

Setup. We examine textual artifacts and language bias through two complementary controls. First, we randomly reassign the test videos over five runs while keeping the trained framework and text candidates unchanged. Second, we evaluate zero-shot Qwen3.5-VL-9B Team (2026) on the same positive–negative choices without video input.

Results. Table 19 reveals residual language bias while also demonstrating a substantial contribution from the correct visual evidence. Randomly reassigning the test videos reduces Overall Acc. of the full framework from 78.98% to 61.17%±1.27% and Comp. Score from 35.78% to 14.28%±1.44%. All four relation categories decline, with the largest reduction on Loc-Replace, from 84.75% to 59.78%±1.19%. Stage-Reorder shows a smaller decrease, from 78.77% to 71.78%±2.53%. The stable above-chance performance across the five shuffled assignments indicates that the text candidates retain predictive cues, while the overall degradation shows that correct video–text correspondence provides substantial additional evidence.

The text-only evaluation makes this bias more explicit. Without video input, Qwen3.5-VL-9B obtains 61.53% Overall Acc. and 15.86% Comp. Score, compared with 74.15% and 33.07% when video is provided. The bias varies substantially across relation types. Stage-Reorder reaches 79.87%, while Dir-Reverse is close to chance at 53.18%. Airport operations and surface movements are highly procedural, with many activities and trajectories following established sequences and prescribed routes. Reordering complete stages can therefore conflict with common operational sequences in airport activities, making the original ordering easier to identify even without the video. Entity-Swap also remains above chance at 64.29%. This may reflect regular associations between entities and their operational roles, since airport vehicles and personnel typically perform particular

Table 19: Analysis of textual artifacts and language bias on AirRelComp. Results for the shuffled pairs report mean ± standard deviation over five random video reassignments.

Setting	Loc-Replace	Dir-Reverse	Stage-Reorder	Entity-Swap	Overall Acc.	Comp. Score
Full framework (similarity comparison)						
Original pairs	84.75	75.51	78.77	70.97	78.98	35.78
Shuffled pairs	59.78±1.19	57.88±2.49	71.78±2.53	57.42±2.91	61.17±1.27	14.28±1.44
Qwen3.5-VL-9B (binary choice)						
Video–text	68.09	69.80	92.47	75.27	74.15	33.07
Text only	58.06	53.18	79.87	64.29	61.53	15.86

services and activities. As a result, some assignments of entities to activities are more plausible than others even without visual evidence. Loc-Replace shows a weaker bias at 58.06%, while movement direction is the least predictable from text alone. Overall, the results show that AirRelComp retains language bias, particularly for temporal ordering and entity participation, although visual input still provides substantial additional evidence across all four relations.

Discussion. The above-chance text-only performance indicates that some controlled relations remain predictable from domain knowledge. This is particularly evident for temporal ordering and entity participation, where airport operations exhibit strong regularities in activity sequences and in the roles performed by different entities. One source of this effect is the current coverage of operational environments: recurring activity sequences, route structures, and entity roles can make some alternatives inherently more predictable. Future extensions can broaden the benchmark across more airports, operational areas, activity types, scene layouts, and combinations of entities and activities. They can also construct harder counterfactuals in which both alternatives remain operationally plausible and differ primarily in their correspondence to the observed video. Greater diversity in camera viewpoints can further increase visual variation and reduce reliance on recurring patterns associated with particular scene layouts. These extensions would reduce such predictable patterns while preserving the diagnostic design of AirRelComp based on controlled relation changes.

C.2 Effect of the Number of Selected Patches and Clusters

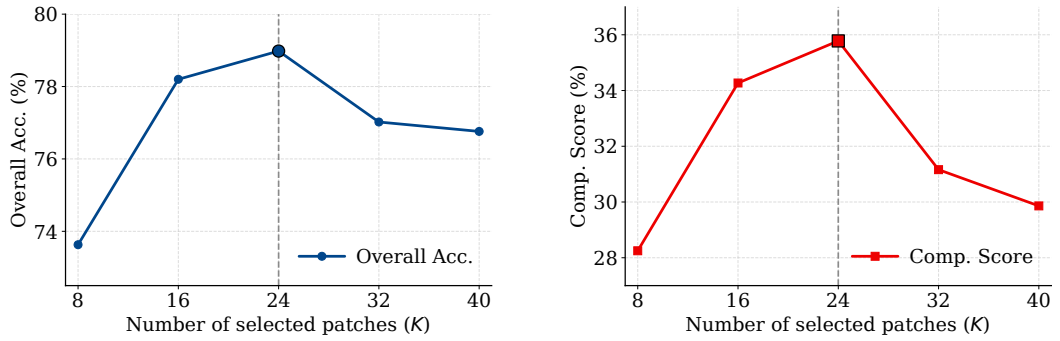
Setup. We analyze the two SpaSelection settings separately while keeping all other configurations unchanged. We first vary the number of selected patches per frame $K \in \{8, 16, 24, 32, 40\}$ with $C = 4$ fixed. We then fix $K = 24$ and vary the number of clusters $C \in \{1, 2, 4, 6, 8, 10\}$. We report Overall Acc. and Comp. Score on AirRelComp.

Results. As shown in Fig. 6, both metrics first improve as K increases and reach their highest values at $K = 24$, before declining with larger K. Overall Acc. increases from 73.63% at $K = 8$ to 78.98% at $K = 24$, while Comp. Score rises from 28.25% to 35.78%. Retaining too few patches may discard useful local evidence from multiple entities and activities, whereas increasing K further can gradually reintroduce redundant or less relevant visual content.

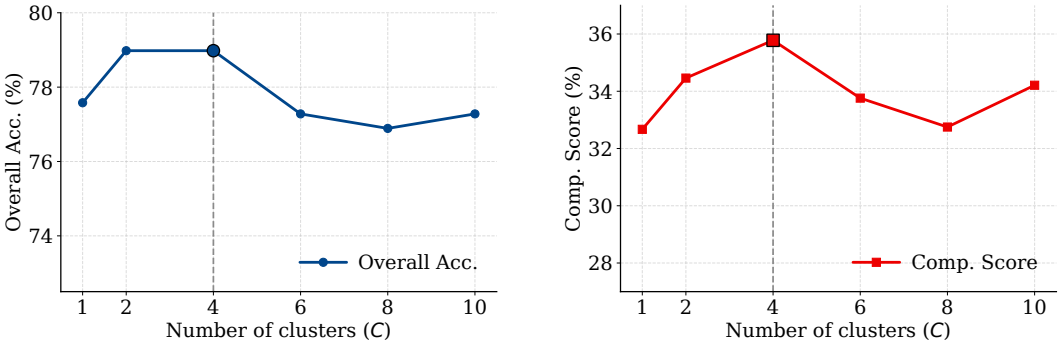
With $K = 24$ fixed, $C = 1$ reduces the selection to score-based Top-K, since only one cluster representative is retained before the remaining positions are filled by score. Increasing C improves performance, with $C = 4$ reaching 78.98% Overall Acc. and the highest Comp. Score of 35.78%. Although $C = 2$ achieves the same Overall Acc., its Comp. Score is lower at 34.46%. Increasing C beyond 4 provides no further improvement. These results support using a small number of clusters to encourage more diverse patch selection while leaving the remaining positions to score-based selection.

C.3 Video–Text Retrieval

Setup. We evaluate the effect of controlled relation learning on standard video–text retrieval. We evaluate T2V and V2T retrieval on the 154 positive validation pairs and report Recall at 1, 5, and 10 in both directions. We compare our full framework with CLIP4Clip Luo et al. (2022), GLSCL Yang et al. (2026), TV-ProxyNet Zhang et al. (2025), and the CLIP-ViP Xue et al. (2023) baseline trained



(a) Overall Acc. (K) (b) Comp. Score (K)



(c) Overall Acc. (C) (d) Comp. Score (C)

Figure 6: Effect of the number of selected patches K and clusters C on AirRelComp.

Table 20: Video–text retrieval performance on the AirRelComp validation set.

Method	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP4Clip Luo et al. (2022)	4.55	20.79	30.52	5.84	17.53	31.17
GLSCL Yang et al. (2026)	7.60	29.10	47.50	6.70	28.20	44.10
TV-ProxyNet Zhang et al. (2025)	14.94	33.12	53.25	7.73	34.02	50.52
CLIP-ViP Xue et al. (2023)	11.69	37.66	59.74	7.14	39.80	55.10
Ours	**16.23**	**44.16**	**63.64**	**8.67**	**41.33**	**57.65**

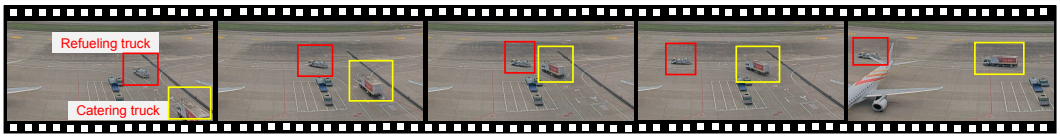
on positive pairs only. As in the main experiments, we select video–text checkpoints by retrieval performance on this validation split.

Results. Table 20 reports the retrieval results. Our full framework achieves the highest recall across all six metrics. The comparison with CLIP-ViP is particularly informative, with T2V R@1 increasing from 11.69% to 16.23% and V2T R@1 from 7.14% to 8.67%. The gains in relation discrimination are accompanied by higher retrieval recall on this validation split.

C.4 Qualitative Analysis

Example. Fig. 7 compares fine-tuned CLIP-ViP with our full framework on an AirRelComp example. The same positive video–text pair is associated with all four types of controlled relation changes. We report the pairwise similarity margin $\Delta = s(v, t^+) - s(v, t^-)$, where a positive margin indicates a correct preference for the positive text.

Analysis. Compared with the fine-tuned CLIP-ViP baseline, our full framework shifts the pairwise similarity margin toward the positive side for all four comparisons. In particular, Loc-Replace and



Positive: On the apron, a **catering truck** **enters** from the **lower** side of the monitored area, afterward **turns** toward the left side of the monitored area; meanwhile, a **refueling truck** exits the monitored area **from right to left**.

Loc-Replace: On the apron, a catering truck enters from the **upper** side of the monitored area, afterward turns toward the left side of the monitored area; meanwhile, a refueling truck exits the monitored area from right to left.

Dir-Reverse: On the apron, a catering truck enters from the lower side of the monitored area, afterward turns toward the left side of the monitored area; meanwhile, a refueling truck exits the monitored area **from left to right.**

Stage-Reorder: On the apron, a catering truck **turns** toward the left side of the monitored area, afterward **enters** from the lower side of the monitored area; meanwhile, a refueling truck exits the monitored area from right to left.

Entity-Swap: On the apron, a **refueling truck** enters from the lower side of the monitored area, afterward turns toward the left side of the monitored area; meanwhile, a **catering truck** exits the monitored area from right to left.

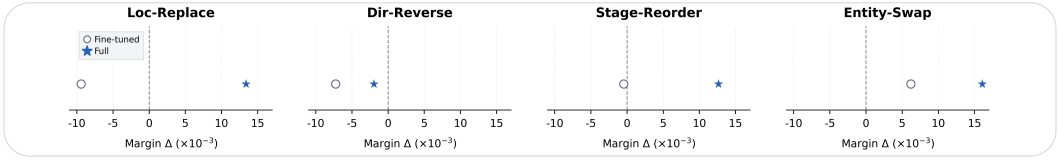


Figure 7: Qualitative comparison of fine-tuned CLIP-ViP and our full framework on an AirRel-Comp example with four controlled relation changes. Colored terms highlight the modified relation elements. The bottom panels report the pairwise similarity margin $\Delta = s(v, t^+) - s(v, t^-)$, where $\Delta > 0$ indicates a correct preference for the positive text.

Stage-Reorder change from incorrect preferences to correct ones, with their margins increasing from -0.0094 to $+0.0134$ and from -0.0005 to $+0.0126$, respectively. Entity-Swap also shows a larger positive separation, while Dir-Reverse remains incorrect but moves closer to the decision boundary, from -0.0073 to -0.0020. This example illustrates improved separation across the four controlled comparisons, including corrected preferences for Loc-Replace and Stage-Reorder.