WHO DOES WHAT, AND WHEN? SYNCHRONIZED CONJUNCTIVE REASONING FOR IMAGE-DISAMBIGUATED VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-Disambiguated Video Temporal Grounding (ID-VTG) aims to localize a temporal segment using a reference image that identifies the target entity and a text query that describes the target event. The task is inherently conjunctive: entity and event conditions must be supported at the same temporal location, while satisfying either condition alone is insufficient. Existing ID-VTG methods effectively integrate image and text cues, yet their entangled proposal relevance leaves temporal entity–event conjunction implicit, limiting their ability to reliably determine when both conditions hold simultaneously. To address this problem, we first introduce Synchronized Conjunctive Grounding (SCG), which factorizes frame-wise entity and event evidence and models their temporal conjunction through a non-compensatory product. The resulting conjunctive evidence is propagated across the temporal pyramid and injected as a signed residual to calibrate proposal relevance, favoring moments jointly supported by both conditions. With entity–event conjunction explicitly modeled, we further characterize its underlying validity structure through Factor-Selective Counterfactual Learning (FCL), making the validity of entity and event factors identifiable. Rather than collapsing intervened queries into generic negatives, FCL exposes the underlying 2×2 factorial validity structure of entity and event conditions and learns a structured four-state energy with unary validity and pairwise interaction potentials. Together, SCG and FCL form a coherent factorize–conjoin–attribute framework for effective entity–event conjunctive grounding. Extensive experiments on existing ID-VTG benchmarks validate the effectiveness of the proposed framework, with consistent improvements across temporal granularities, ambiguity settings, and cross-domain generalization scenarios.

1 INTRODUCTION

Video Temporal Grounding (VTG) aims to localize the temporal segment described by a natural-language query in an untrimmed video. Existing methods broadly span proposal-based matching and regression (Gao et al., 2017; Zhang et al., 2020), query-conditioned detection and metric alignment (Lei et al., 2021; Wang et al., 2022; Moon et al., 2023; Xiao et al., 2024), and more recent scalable (Mu et al., 2024), pretrained (Luo et al., 2023), or MLLM-based grounding (Wu et al., 2025; Zhang et al., 2026). Despite this progress, text alone can be ambiguous when multiple entities perform similar events. The recently introduced Image-Disambiguated Video Temporal Grounding (ID-VTG) (Zheng et al., 2026a) addresses this ambiguity by pairing the event description with a reference image that identifies the target entity. The task therefore requires determining *who* performs *what*, and critically, *when* both conditions hold simultaneously.

VGD-Agg (Zheng et al., 2026a) provides a strong foundation by combining fine-grained frame–reference matching with text-guided temporal grounding. However, image and text cues ultimately contribute to entangled proposal relevance, leaving their temporal conjunction implicit. Consequently, a proposal may receive high relevance when the target entity performs another event, when a distractor performs the queried event, or when strong entity and event responses arise at different temporal locations. We refer to this insufficient modeling of simultaneous entity–event support as the *temporal co-occurrence gap*.

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.

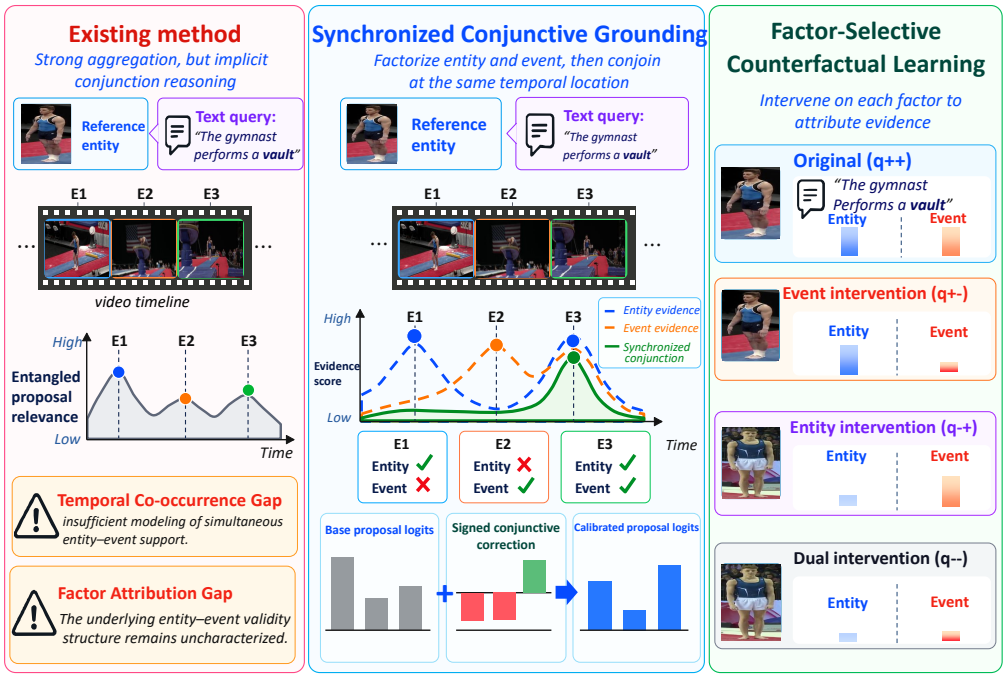


Figure 1: Existing ID-VTG models leave entity–event co-occurrence implicit in entangled proposal relevance, allowing partial matches to rank highly while leaving the constituent validity factors unresolved. Our framework explicitly models synchronized entity–event support and identifies their validity through selective counterfactual reasoning.

To address this gap, we first introduce **Synchronized Conjunctive Grounding (SCG)**. SCG factorizes frame-wise entity and event evidence and couples them pointwise on the finest shared temporal grid through a non-compensatory product, ensuring that strong evidence from one factor cannot compensate for weak evidence from the other. The resulting synchronized signal is propagated across the temporal pyramid and injected as a signed residual into the base proposal logits, explicitly calibrating proposal relevance with same-time conjunctive support.

With entity–event conjunction explicitly modeled, we further characterize its underlying validity structure. Factual conjunction supervision alone cannot uniquely determine whether the entity, the event, or both are valid, leaving the constituent factor validities unidentifiable. We refer to this ambiguity as the *factor attribution gap*. To expose this structure and resolve the resulting gap, we introduce **Factor-Selective Counterfactual Learning (FCL)**, which selectively intervenes on the entity and event queries. These interventions instantiate a 2×2 validity space covering joint validity, two single-factor violations, and joint invalidity. Rather than collapsing these configurations into generic negatives, FCL learns a structured four-state energy directly from the SCG evidence using unary factor-validity and pairwise interaction potentials. FCL is used only during training and introduces no additional inference head.

SCG and FCL form a coherent *factorize–conjoin–attribute* progression: entity and event evidence is first made explicit, synchronized to determine where both conditions hold, and then counterfactually supervised to make the validity factors identifiable and attributable. Our contributions are:

- We propose Synchronized Conjunctive Grounding to address the temporal co-occurrence gap by explicitly modeling synchronized entity–event conjunction for proposal calibration.
- We reveal the factorial validity structure underlying conjunctive grounding and introduce Factor-Selective Counterfactual Learning to resolve its factor attribution gap through selective interventions and structured four-state supervision.
- Extensive experiments on three benchmarks demonstrate substantial improvements across temporal granularities, ambiguity types, and out-of-domain settings, with controlled analyses further validating the proposed designs.

2 Related Work

Video temporal grounding. Early VTG methods formulate temporal localization through proposal matching, boundary regression, and cross-modal interaction (Gao et al., 2017; Zhang et al., 2020; Soldan et al., 2021). Subsequent work develops direct set prediction, metric learning, query-dependent representations, and unified grounding objectives (Lei et al., 2021; Wang et al., 2022; Moon et al., 2023; 2026b; Lin et al., 2023; Xiao et al., 2024). Recent studies further target scalability and generalization through late fusion and pretrained vision–language priors (Mu et al., 2024; Luo et al., 2023; 2024; Zheng et al., 2024), as well as more discriminative, context-robust, or hierarchical temporal representations (Sun et al., 2025; Zhou et al., 2025; An & Grauman, 2026; Moon et al., 2026a). In parallel, MLLM-based approaches explore explicit temporal encoding, query enrichment, stronger training recipes, and open-world supervision (Wu et al., 2025; Pramanick et al., 2025; Zhang et al., 2026; Zheng et al., 2026b). These advances substantially improve text-conditioned temporal localization, whereas ID-VTG additionally requires the localized event to be associated with a reference-specified entity at the same temporal location.

Multimodal query grounding. Visual queries provide instance-specific appearance cues that can be difficult to express precisely in language. MINOTAUR (Goyal et al., 2023) unifies video grounding across visual, textual, and activity queries, while ICQ (Zhang et al., 2025) studies event localization conditioned on multimodal image–text queries. ID-VTG (Zheng et al., 2026a) focuses specifically on referential ambiguity, using an image to identify the target entity and text to describe the target event. Its VGD-Agg framework performs fine-grained frame–reference disambiguation before integrating the resulting representation into temporal grounding. Our work addresses a complementary issue: rather than introducing another multimodal fusion mechanism, we explicitly model whether entity and event evidence are supported at the same temporal location and use this conjunctive evidence to calibrate proposal relevance.

Negative and counterfactual learning. Negative learning has been explored in temporal grounding to improve fine-grained video-language discrimination (Wang et al., 2022; Zheng et al., 2022a;b; Moon et al., 2023). Causal and counterfactual methods exploit interventions to reduce spurious correlations, improve robustness, or strengthen factual–counterfactual discrimination (Nan et al., 2021; Yang et al., 2021; Jiang et al., 2024; Xu et al., 2025; Wang et al., 2025). These methods primarily use interventional or counterfactual signals for de-biasing, robustness, or relevance discrimination. FCL instead uses counterfactual intervention to reveal the factorial structure underlying conjunctive grounding: entity validity and event validity constitute two semantically distinct factors whose combinations define a structured grounding state space. It therefore attributes conjunctive outcomes to their constituent factors, rather than treating different counterfactuals as interchangeable negatives.

3 Method

3.1 Overview

Given an untrimmed video V, a reference image I, and an event description x, ID-VTG predicts a temporal interval $\hat{y} = (\hat{t}_s, \hat{t}_e)$ in which the reference-specified entity performs the described event. During training, the corresponding ground-truth interval is denoted by $y = (t_s, t_e)$. We adopt VGD-Agg (Zheng et al., 2026a) as the base framework and omit the batch dimension for clarity. As shown in Fig. 2, SCG constructs aligned entity and event evidence on the finest temporal grid, forms their explicit conjunction, and uses it to calibrate the base proposal logits across the temporal pyramid. FCL further supervises the constituent evidence during training through factor-selective interventions. Following the notation of VGD-Agg, let $\mathbf{v} \in \mathbb{R}^{T \times d}$ denote the frame-level video features, where T is the valid sequence length. We denote the encoded text tokens by $\mathbf{Q} \in \mathbb{R}^{N_q \times d}$ and the projected reference image representation by $\mathbf{R} \in \mathbb{R}^d$. Let $\{\mathbf{Z}^\ell\}_{\ell=0}^{L_p-1}$ denote the query-independent temporal feature pyramid produced by the Fast Branch, where $\mathbf{Z}^\ell \in \mathbb{R}^{T_\ell \times d}$ and L_p is the number of pyramid levels. The finest level shares the temporal resolution of $\mathbf{v}$, with $T_0 = T$. Each index t on this finest grid corresponds to a sampled frame, and we therefore refer to the resulting scores as frame-wise evidence. Before SCG calibration, the base model predicts proposal classification logits $\mathbf{a}^\ell \in \mathbb{R}^{T_\ell}$ and boundary offsets $\mathbf{o}^\ell \in \mathbb{R}^{T_\ell \times 2}$ at each pyramid level. We refer to $\mathbf{a}^\ell$ as the base proposal logits.

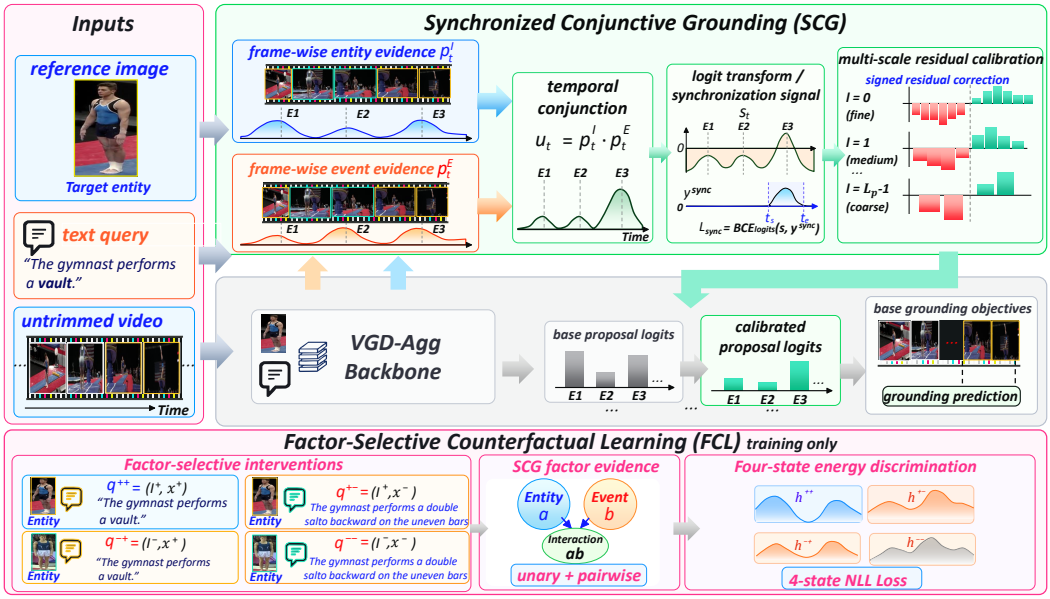


Figure 2: Our method follows a *factorize–conjoin–attribute* design and is built upon the VGD-Agg backbone. Synchronized Conjunctive Grounding (SCG) first factorizes frame-wise entity and event evidence, establishes their same-time conjunction with a non-compensatory product, and propagates the resulting conjunctive evidence across proposal scales for residual calibration. Building on SCG, Factor-Selective Counterfactual Learning (FCL) further makes entity and event validity identifiable through selective counterfactual interventions and a structured four-state energy model. SCG and FCL are jointly optimized in a unified framework, integrating explicit temporal conjunction with factor-specific supervision for the ID-VTG task.

3.2 Synchronized Conjunctive Grounding

Factorized entity and event evidence. The reference image identifies the target entity, whereas the text describes which event should occur. We introduce separate entity and event scorers:

$$\mathbf{e}^I = g_I(\mathbf{v}, \mathbf{R}), \qquad \mathbf{e}^E = g_E(\mathbf{Z}^0, \mathbf{Q}), \tag{1}$$

where both outputs belong to $\mathbb{R}^T$. The two scorers share the same architecture but have independent parameters. Each consists of a query-conditioned Transformer decoder, layer normalization, temporal convolution, ReLU, and a scalar prediction head. The frame-level representation $\mathbf{v}$ preserves fine-grained appearance information for entity discrimination, whereas the finest pyramid features $\mathbf{Z}^0$ provide temporally contextualized representations for event recognition. Since $T_0 = T$, the two evidence sequences are defined on the same temporal grid. We convert them to bounded evidence:

$$p_t^I = \sigma(e_t^I), \qquad p_t^E = \sigma(e_t^E), \tag{2}$$

Explicit temporal conjunction. On the finest shared temporal grid, we combine the two factor-evidence sequences pointwise:

$$\mathbf{u} = \mathbf{p}^I \odot \mathbf{p}^E, \qquad u_t = p_t^I p_t^E \le \min\{p_t^I, p_t^E\}, \quad t = 1, \dots, T. \tag{3}$$

The product is non-compensatory: strong evidence from one factor cannot offset weak evidence from the other, directly reflecting the requirement that both conditions be supported. We compute the conjunction on the finest shared temporal grid before cross-scale resampling. This fine-to-coarse ordering preserves temporal correspondence and prevents entity and event responses from different fine-grained locations from being spuriously combined after coarsening. To obtain a signed signal compatible with the original proposal logits, we map the conjunction into logit space, where the small number ϵ ensures numerical stability.

$$s_t = \log \frac{\widehat{u}_t}{1 - \widehat{u}_t}, \qquad \widehat{u}_t = \mathrm{clip}(u_t, \epsilon, 1 - \epsilon), \tag{4}$$

Synchronized proposal calibration and supervision. We construct the synchronized signal on the finest shared temporal grid, where entity and event evidence retain their pointwise temporal correspondence. The fine-grained signal is then propagated across the temporal pyramid through interpolation and injected into the corresponding base proposal logits:

$$\widetilde{a}_j^{\ell} = a_j^{\ell} + \lambda \left[\mathcal{I}_{T \to T_\ell}(\mathbf{s})\right]_j, \qquad \ell = 0, \ldots, L_p - 1, \tag{5}$$

where j indexes temporal locations at pyramid level ℓ, $\mathcal{I}_{T \to T_\ell}$ denotes temporal interpolation from the finest grid to the resolution of level ℓ, and λ is a learned scalar. Constructing the conjunction before pyramid coarsening preserves fine-grained temporal correspondence and prevents entity and event responses from different locations from being spuriously associated at coarser resolutions. The signed residual reinforces jointly supported proposals while suppressing partial matches.

We directly supervise the synchronized evidence on the finest shared temporal grid using a truncated Gaussian target y_t^{sync}, centered at the midpoint of the ground-truth interval and set to zero outside the interval. Its width is defined by $\sigma_y = \max(\rho(t_e - t_s), 1)$, with $\rho = 0.25$. Using binary cross-entropy with logits, the synchronized supervision objective is

$$\mathcal{L}_{\text{sync}} = \frac{1}{T} \sum_{t=1}^{T} \ell_{\text{BCEWithLogits}}\left(s_t, y_t^{\text{sync}}\right). \tag{6}$$

3.3 Factor-Selective Counterfactual Learning

Although SCG makes entity and event evidence explicit, conjunctive supervision alone cannot determine whether a failed conjunction arises from the entity factor, the event factor, or both. We therefore introduce FCL, which selectively intervenes on the two query factors and supervises their factor-specific validity. Rather than treating all counterfactuals as interchangeable negatives, FCL constructs controlled and challenging validity states and models both individual factors and their interaction.

Factor-selective counterfactual construction. For a factual sample (V, I^+, x^+, y), we keep the video and annotated interval fixed and selectively intervene on the entity and event factors:

$$\begin{aligned} q^{++} &= (I^+, x^+), \quad q^{+-} = (I^+, x^-), \\ q^{-+} &= (I^-, x^+), \quad q^{--} = (I^-, x^-). \end{aligned} \tag{7}$$

These correspond to the factual, event-intervened, entity-intervened, and dual-intervened inputs, respectively. All states retain the original video V and the same temporal support y for state supervision. To obtain diverse and challenging counterfactuals, replacement candidates are drawn exclusively from the training set, using both same-video and cross-video sources. Potential counterfactual noise is automatically reduced using separate eligibility criteria for entity and event candidates, together with similarity screening and duplicate or near-duplicate removal. For each factual training sample, the remaining entity and event candidates are separately ranked by their respective similarity measures to form two sample-specific top-K pools offline, from which one replacement per factor is randomly sampled during training. The dual intervention is constructed when valid replacements for both factors are available. Detailed settings are given in the implementation details.

Counterfactual state representation. Each input q^z is associated with a validity state $z = (r_I, r_E)$, where $r_I, r_E \in \{-1, +1\}$ indicate entity and event validity. We consider $\mathcal{Z} = \{z^{++}, z^{+-}, z^{-+}, z^{--}\}$, corresponding to $(+1, +1)$, $(+1, -1)$, $(-1, +1)$, and $(-1, -1)$, respectively. The mixed states isolate single-factor violations, while z^{--} represents a distinct joint-invalid configuration. For an input q^z, we convert the SCG logits into centered factor evidence:

$$a_{t,z} = 2\sigma(e_{t,z}^I) - 1, \qquad b_{t,z} = 2\sigma(e_{t,z}^E) - 1. \tag{8}$$

Positive and negative values indicate evidence for factor validity and violation, respectively.

Structured four-state discrimination. For each constructed input, we define four compatibility logits, ordered as $(++, +-, -+, --)$:

$$\mathbf{h}_{t,z} = \frac{1}{T_{\text{cf}}} \begin{bmatrix} a_{t,z} + b_{t,z} + \eta a_{t,z} b_{t,z} \\ a_{t,z} - b_{t,z} - \eta a_{t,z} b_{t,z} \\ -a_{t,z} + b_{t,z} - \eta a_{t,z} b_{t,z} \\ -a_{t,z} - b_{t,z} + \eta a_{t,z} b_{t,z} \end{bmatrix}. \tag{9}$$

Here, $T_{\text{cf}} > 0$ is a temperature and η controls factor interaction. The unary terms act as first-order potentials for entity and event validity, while the bilinear term $\eta a_{t,z} b_{t,z}$ provides a second-order interaction potential. The interaction separates same-validity $(++, --)$ from mixed-validity $(+-, -+)$ configurations, while the unary terms distinguish the states within each pair. Together, they induce structured four-state discrimination directly from the SCG evidence without introducing a separate prediction head. Let $\mathcal{G} = \{t : t \in y\}$ denote the temporal positions inside the factual interval. We uniformly pool the compatibility logits:

$$\overline{h}_z^{z'} = \frac{1}{|\mathcal{G}|} \sum_{t \in \mathcal{G}} h_{t,z}^{z'}. \tag{10}$$

The interaction $a_{t,z} b_{t,z}$ is computed before pooling, preserving its dependence on same-time factor evidence. The pooled logits are normalized as

$$P_\theta(z' \mid q^z, V, y) = \frac{\exp(\overline{h}_z^{z'})}{\sum_{\tilde{z} \in \mathcal{Z}} \exp(\overline{h}_z^{\tilde{z}})}. \tag{11}$$

Equivalently, $E_z(z') = -\overline{h}_z^{z'}$ defines a structured energy over the four validity states, composed of first-order factor potentials and a second-order interaction. Let $\mathcal{S}_i \subseteq \mathcal{Z}$ denote the available states used by FCL for sample i. We optimize over $\mathcal{B}_{\text{cf}}$, the set of training samples with non-empty $\mathcal{S}_i$:

$$\mathcal{L}_{\text{FCL}} = \frac{1}{|\mathcal{B}_{\text{cf}}|} \sum_{i \in \mathcal{B}_{\text{cf}}} \frac{1}{|\mathcal{S}_i|} \sum_{z \in \mathcal{S}_i} -\log P_\theta(z \mid q_i^z, V_i, y_i), \tag{12}$$

3.4 Training Objective and Inference

We retain the base proposal classification, regression, and ranking losses of VGD-Agg:

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(\widetilde{\mathbf{a}}^\ell) + 2\mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{rank}} + \lambda_{\text{sync}} \mathcal{L}_{\text{sync}} + \lambda_{\text{cf}} \mathcal{L}_{\text{FCL}}. \tag{13}$$

We set $\lambda_{\text{sync}} = \lambda_{\text{cf}} = 0.5$, $T_{\text{cf}} = 1$, and $\eta = 0.5$. For the factual query, $\mathcal{L}_{\text{cls}}$ is applied to the calibrated proposal logits $\widetilde{\mathbf{a}}^\ell$, allowing the grounding objective to jointly optimize the base grounding pathway and the SCG evidence branches. The regression and ranking losses retain their original formulations, while intervened inputs contribute only to $\mathcal{L}_{\text{FCL}}$. All components are jointly optimized. At inference, SCG calibrates the base proposal logits using Eq. 5. FCL, including counterfactual construction and state supervision, is used only during training.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate on the existing ID-VTG benchmarks using their official splits (Zheng et al., 2026a). **IDVTG-Gym** contains approximately 14.7K queries over 204.1 hours of FineGym videos, with hierarchical annotations ranging from event-level categories (e.g., uneven bars) to atomic sub-actions (e.g., giant circle). It therefore evaluates grounding at both fine-grained sub-action and holistic-event scales. **IDVTG-InternVid** contains approximately 62.1K queries over 302.7 hours of more diverse entities, events, and visual contexts. **Web** consists of independently collected videos and serves as an out-of-domain evaluation. We report top-1 recall at temporal IoU thresholds 0.5 and 0.7 ($R^1_{.5}$ and $R^1_{.7}$) and mean IoU (mIoU).

Implementation. We build on the released VGD-Agg implementation and use same offline CLIP ViT-L/14 features (Radford et al., 2021), text processing, and temporal decoding. Experiments use 1 NVIDIA A800 GPU, a batch size of 4, and AdamW (Loshchilov & Hutter, 2019) with a learning rate of 2e-4. Models are trained for 50 epochs on IDVTG-Gym and 20 epochs on IDVTG-InternVid. For FCL, entity and event candidates are ranked by visual and textual similarity, respectively. For each factual training instance, we retain top-K eligible candidates for each factor and randomly sample one replacement from the corresponding pool on the fly during training, with $K = 20$. Entity candidates are restricted to visual similarity in $[0.5, 0.995]$, while event similarity is capped at 0.8 to avoid paraphrastic replacements; same-video candidates are further required to satisfy $\text{tIoU} \leq 0.1$. We use VGD-Agg as our primary baseline and report both its reported and reproduced results. Additional experimental analyses and theoretical discussions are provided in the appendix.

Table 1: Comparison on IDVTG-Gym dataset across temporal granularities. Our SCG and SCG+FCL models are built upon the reproduced VGD-Agg† baseline.

Method	Overall			Sub-action			Holistic Event		
	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU
RaTSG (Dong et al., 2024)	18.45	15.99	16.96	6.64	3.98	5.54	26.28	23.50	27.64
UVCOM (Xiao et al., 2024)	18.07	15.00	17.32	5.20	2.43	6.86	25.07	22.54	22.99
CG-DETR (Moon et al., 2026b)	27.96	23.36	26.99	6.64	2.77	10.14	44.47	39.86	40.29
ICQ (Zhang et al., 2025)	46.12	42.24	40.24	43.92	35.73	34.10	47.87	47.17	45.06
SnAG (Mu et al., 2024)	53.64	48.82	46.73	51.66	41.37	40.32	54.83	54.05	51.46
VGD-Agg (Zheng et al., 2026a) (reported)	61.83	56.53	54.17	55.53	44.91	43.29	66.58	65.36	62.53
VGD-Agg† (reproduced)	62.96	58.33	55.07	56.31	46.57	43.62	67.36	66.58	63.31
Ours (SCG)	70.62	64.76	61.80	63.50	51.03	49.38	76.68	75.46	71.91
Ours (SCG+FCL)	**73.56**	**67.08**	**64.33**	**64.27**	**51.11**	**49.73**	**80.77**	**79.46**	**75.70**

Table 2: Comparison on IDVTG-InternVid and Web (OOD) datasets.

Method	InternVid			Web (OOD)		
	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU
RaTSG (Dong et al., 2024)	30.58	19.71	30.91	7.96	5.79	8.53
UVCOM (Xiao et al., 2024)	30.11	18.89	30.07	7.70	5.62	8.10
CG-DETR (Moon et al., 2026b)	44.81	31.62	43.80	13.86	5.91	–
ICQ (Zhang et al., 2025)	31.04	25.39	29.90	7.84	4.53	8.31
SnAG (Mu et al., 2024)	44.26	35.46	41.70	16.48	9.06	16.53
VGD-Agg (Zheng et al., 2026a) (reported)	51.21	41.24	48.30	21.99	13.25	22.34
VGD-Agg† (reproduced)	52.15	41.89	49.45	24.82	14.08	24.69
Ours (SCG)	55.30	43.81	52.09	27.69	15.87	27.34
Ours (SCG+FCL)	**57.52**	**46.34**	**53.79**	**29.69**	**16.58**	**29.04**

4.2 Comparison with Existing Methods

IDVTG-Gym. Table 1 shows that our complete model improves the reproduced VGD-Agg baseline by 10.60, 8.75, and 9.26 points in Overall $R^1_{.5}$, $R^1_{.7}$, and mIoU, respectively. Both SCG and FCL contribute to the overall performance improvement, with FCL providing further gains over SCG. Consistent gains are observed across both temporal granularities: fine-grained Sub-actions characterized by short, atomic movements, and Holistic Events requiring longer-range event dependencies, with mIoU improvements of 6.11 and 12.39 points, respectively. The larger gain on Holistic Events may arise because longer events contain more partial matches, making synchronized entity–event reasoning particularly useful.

IDVTG-InternVid and Web. Table 2 shows that our complete model improves the reproduced VGD-Agg baseline by 5.37, 4.45, and 4.34 points on InternVid in $R^1_{.5}$, $R^1_{.7}$, and mIoU, respectively. SCG improves performance on both InternVid and Web, while FCL provides further gains across all three metrics. These improvements persist across the diverse entities, events, and visual contexts of InternVid. We directly evaluate the models trained on IDVTG-InternVid on Web without target-domain adaptation, where our complete model improves over the reproduced baseline by 4.87, 2.50, and 4.35 points. The consistent improvement under this domain shift suggests that the learned entity–event reasoning transfers beyond the source-domain distribution.

Analysis under Different Ambiguity Types. Table 3 reveals complementary effects of SCG and FCL across different ambiguity types. On *Text Amb. Only*, where event semantics are insufficient to identify the target entity, FCL further improves SCG by 5.45, 5.33, and 4.13 points in $R^1_{.5}$, $R^1_{.7}$, and mIoU, respectively. On *Img Amb. Only*, where visual identity is insufficient to determine the target event, SCG already provides a clear gain through explicit same-time conjunction. For *Full Amb.*, where neither factor alone is sufficient, SCG accounts for most of the improvement while FCL provides further refinement.

Table 3: Performance on the official IDVTG-InternVid ambiguity subsets.

Method	Full Amb.			Text Amb. Only			Img Amb. Only		
	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU
VGD-Agg†	38.28	28.22	38.02	53.58	44.06	50.99	43.49	31.96	41.58
Ours (SCG)	40.66	30.71	40.28	57.03	47.09	54.20	46.47	33.20	44.42
Ours (SCG+FCL)	**41.60**	**31.22**	**40.90**	**62.48**	**52.42**	**58.33**	**46.58**	**35.00**	**45.07**

Table 4: Mechanistic analysis of SCG on IDVTG-Gym and IDVTG-InternVid.

Method	Gym Overall			InternVid Overall		
	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU
VGD-Agg† (baseline)	62.96	58.33	55.07	52.15	41.89	49.45
Coupling operator						
Mean	60.17	55.63	52.69	54.46	43.71	51.67
Minimum	64.90	59.74	56.91	54.32	43.24	51.28
Product	**70.62**	**64.76**	**61.80**	**55.30**	**43.81**	**52.09**
Synchronized supervision and residual calibration						
Product w/o $\mathcal{L}_{\text{sync}}$	66.60	61.31	58.39	53.43	43.52	50.49
Product, inference $\lambda = 0$	67.08	61.59	58.76	54.11	42.90	51.23
Full SCG	**70.62**	**64.76**	**61.80**	**55.30**	**43.81**	**52.09**

4.3 Ablation Study

The product coupling is important. Table 4 shows that Product consistently outperforms Mean and Minimum on both datasets. Mean performs poorly because it is compensatory: a strong response from one factor can offset weak support from the other, allowing partial matches to retain relatively high conjunctive evidence. Minimum avoids this failure by requiring both factors to be strong, but its output is determined entirely by the weaker response and therefore discards confidence carried by the other factor. Product is both non-compensatory and jointly responsive: it penalizes mismatched factor responses while remaining sensitive to the confidence of both.

Both synchronized supervision and calibration are important. As shown in Table 4, removing either $\mathcal{L}_{\text{sync}}$ or residual calibration (inference) consistently degrades performance, confirming the complementary benefits of fine-grid conjunctive supervision and its explicit injection into proposal logits. The remaining gain with $\lambda = 0$ further suggests that synchronized supervision also improves the learned grounding representation beyond its inference-time calibration effect.

Structured factor modeling goes beyond counterfactual augmentation. Table 5 shows that the gains are not merely due to introducing additional counterfactual samples. Using the same counterfactual set, binary supervision already improves over SCG, while factor-wise supervision further benefits from preserving entity and event validity. With $\eta = 0$, the two factors are modeled independently; restoring their interaction yields the full four-state energy model and the best performance, demonstrating the benefit of structured factor modeling.

Our counterfactual construction strategy provides more informative supervision. As shown in Table 5, our strategy benefits from combining same- and cross-video candidates, while outperforming random and similarity-hard alternatives by reducing noisy counterfactuals through deduplication, and similarity/temporal constraints. We further compare with a textual counterfactual construction strategy using Qwen3-8B (Yang et al., 2025). Such generic text generation does not explicitly exploit the event ambiguity distribution and incurs higher construction cost.

Generality analysis. We further integrate SCG and FCL into the SnAG method. Compared with the reported SnAG results, our method consistently improves performance on Gym, InternVid, and Web, providing additional evidence that our design transfers beyond the VGD-Agg framework.

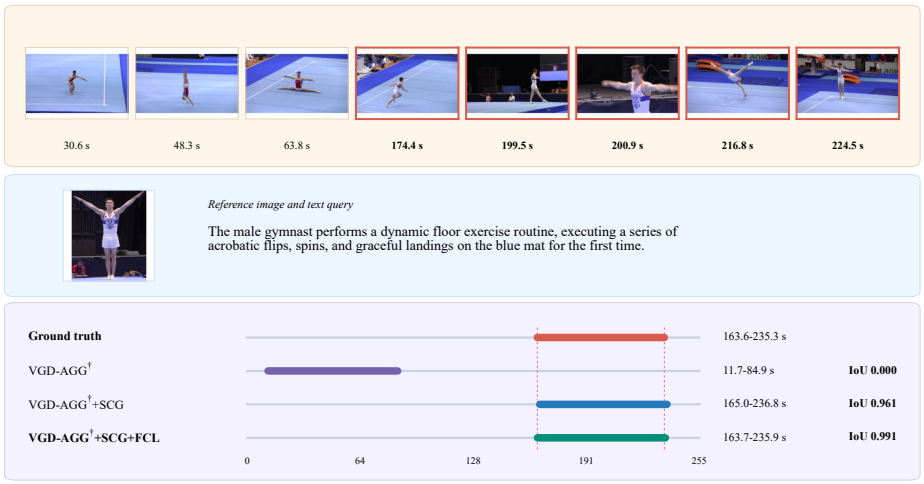


Figure 3: Qualitative temporal grounding comparison on IDVTG-Gym.

Table 5: Analysis of factor-selective counterfactual learning (FCL) on IDVTG-InternVid.

Method	$R^1_{.5}$	$R^1_{.7}$	mIoU
FCL beyond generic counterfactual augmentation			
SCG	55.30	43.81	52.09
SCG + Binary counterfactual supervision	56.27	45.11	52.95
SCG + Factor-wise supervision ($\eta = 0$)	56.59	45.75	53.29
Structured four-state FCL	**57.52**	**46.34**	**53.79**
Counterfactual sample construction (fixed FCL objective)			
Random	55.75	44.78	52.88
Similarity-hard	56.44	44.81	53.03
Same-video	55.50	44.30	52.61
Cross-video	57.01	45.40	53.50
Textual counterfactual construction (Qwen3-8B)	57.16	45.76	53.58
FCL strategy	**57.52**	**46.34**	**53.79**

Table 6: Generality of our framework across other grounding methods on IDVTG benchmarks.

	IDVTG-Gym			IDVTG-InternVid			Web		
Method	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU	$R^1_{.5}$	$R^1_{.7}$	mIoU
SnAG (Mu et al., 2024) (reported)	53.64	48.82	46.73	44.26	35.46	41.70	16.48	9.06	16.53
SnAG + SCG + FCL	**69.21**	**64.24**	**60.67**	**52.23**	**42.34**	**48.87**	**23.50**	**13.36**	**22.61**

Qualitative Analysis. Figure 3 presents a representative example on IDVTG-Gym. VGD-Agg† localizes an earlier, visually similar floor-exercise segment performed by another gymnast. SCG shifts the prediction to the interval jointly supported by the reference gymnast and queried event. The model trained with FCL further refines the temporal boundaries, achieving more precise localization.

5 Conclusion

In this paper, we study the recently introduced ID-VTG task, which requires identifying the reference-specified entity, recognizing the queried event, and determining when both conditions hold simultaneously. To address this conjunctive challenge, we develop a unified *factorize–conjoin–attribute* framework that disentangles frame-wise entity and event evidence, establishes their synchronized conjunction for proposal calibration, and further attributes factor validity through structured counterfactual supervision. Extensive experiments across ID-VTG evaluation settings demonstrate the effectiveness and superiority of the proposed framework.

AI USE STATEMENT

The authors used a large language model (LLM) to refine the language and improve the clarity and readability of the manuscript, including sentence rephrasing and grammar checking. In addition, Qwen3-8B was used solely to generate textual counterfactual samples for a controlled comparison of counterfactual construction strategies. It was not used in the proposed retrieval-based construction strategy or to formulate the research ideas, design the proposed methodology or experiments, or analyze the results. All research concepts, methodological decisions, experimental protocols, and analyses were developed and conducted by the authors. The authors take full responsibility for the content of the manuscript.

REFERENCES

Joungbin An and Kristen Grauman. Hieramamba: Video temporal grounding via hierarchical anchor-mamba pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16954–16965, 2026.

Jianfeng Dong, Xiaoman Peng, Daizong Liu, Xiaoye Qu, Xun Yang, Cuizhu Bao, and Meng Wang. Temporal sentence grounding with relevance feedback in videos. In *Advances in Neural Information Processing Systems*, volume 37, pp. 43107–43132, 2024.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285. IEEE, 2017.

Raghav Goyal, Effrosyni Mavroudi, Xitong Yang, Sainbayar Sukhbaatar, Leonid Sigal, Matt Feiszli, Lorenzo Torresani, and Du Tran. Minotaur: Multi-task video grounding from multimodal queries. *arXiv preprint arXiv:2302.08063*, 2023.

Xun Jiang, Zhuoyuan Wei, Shenshen Li, Xing Xu, Jingkuan Song, and Heng Tao Shen. Counterfactually augmented event matching for de-biased temporal sentence grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6472–6481, 2024.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11846–11858, 2021.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023.

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

Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23045–23055, 2023.

Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Zero-shot video moment retrieval from frozen vision-language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5452–5461. IEEE, 2024.

Sungho Moon, Seunghun Lee, Jiwan Seo, and Sunghoon Im. Cva: Context-aware video-text alignment for video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17578–17587, 2026a.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023–23033, 2023.

WonJun Moon, Sangeek Hyun, Subeen Lee, and Jae-Pil Heo. Correlation-guided calibration of query dependency for video temporal grounding. *Pattern Recognition*, 174:112984, 2026b.

Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18930–18940, 2024.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2765–2775, 2021.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24297–24308, 2025.

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

Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3224–3234, 2021.

Xiaolong Sun, Le Wang, Sanping Zhou, Liushuai Shi, Kun Xia, Mengnan Liu, Yabing Wang, and Gang Hua. Moment quantization for video temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20137–20146, 2025.

Qiyi Wang, Senda Chen, and Ying Shen. CausalVTG: Towards robust video temporal grounding via causal inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3461.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2613–2623, 2022.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13754–13765, 2025.

Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18709–18719, 2024.

Yenan Xu, Wanru Xu, and Zhenjiang Miao. Counterfactual contrastive learning for weakly supervised temporal sentence grounding. *Neurocomputing*, 624:129508, 2025. doi: 10.1016/j.neucom.2025.129508.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1–10, 2021. doi: 10.1145/3404835.3462823.

Gengyuan Zhang, Mang Ling Ada Fok, Jialu Ma, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. Localizing events in videos with multimodal queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3339–3351, 2025.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12870–12877, 2020.

Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3517–3525, 2022a.

Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15534–15543, 2022b.

Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, pp. 20–37, 2024.

Minghang Zheng, Jingli Wei, Hongyi Yang, and Yang Liu. Id-vtg: Image-disambiguated video temporal grounding. *arXiv preprint arXiv:2608.20127*, 2026a.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. Omnivtg: A large-scale dataset and training paradigm for open-world video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24620–24629, 2026b.

Xinyang Zhou, Fanyue Wei, Lixin Duan, Angela Yao, and Wen Li. The devil is in the spurious correlations: Boosting moment retrieval with dynamic learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20981–20990, 2025.

APPENDIX

A THEORETICAL ANALYSIS OF SCG AND FCL

We provide a theoretical analysis of the two key structural designs in our framework. We first examine why the product operator used by SCG is better suited to conjunctive grounding than Mean or Minimum. We then show that conjunctive supervision alone leaves the constituent factors underdetermined, which motivates the structured factor supervision introduced by FCL.

A.1 ANALYSIS OF SYNCHRONIZED CONJUNCTIVE GROUNDING

Let $p = \sigma(e^I)$ and $q = \sigma(e^E)$ denote the entity and event evidence at one temporal location. For a coupling operator $u = \phi(p, q)$, ignoring the numerical clipping used only for stability, the synchronized BCE objective can be written as

$$\ell(u; y) = -y \log u - (1 - y) \log(1 - u), \qquad y \in [0, 1]. \tag{14}$$

This formulation directly includes the soft synchronization target used in practice; the cases $y = 1$ and $y = 0$ below expose its positive and negative optimization behavior most clearly.

Product conjunction and its gradients. For SCG, $u_{\text{prod}} = pq$. Using $\partial p/\partial e^I = p(1 - p)$ and $\partial q/\partial e^E = q(1 - q)$, we obtain

$$\frac{\partial \ell_{\text{prod}}}{\partial e^I} = \frac{(pq - y)(1 - p)}{1 - pq}, \qquad \frac{\partial \ell_{\text{prod}}}{\partial e^E} = \frac{(pq - y)(1 - q)}{1 - pq}. \tag{15}$$

Hence the product confidence pq is driven toward the synchronized target y. In particular, Eq. 15 remains valid for the Gaussian-shaped soft target used around temporal boundaries.

For a positive conjunctive target ($y = 1$),

$$\ell^+_{\text{prod}} = -\log(pq) = -\log p - \log q, \qquad \nabla \ell^+_{\text{prod}} = \big(-(1 - p), -(1 - q)\big). \tag{16}$$

Thus, Product decomposes the positive conjunctive objective into two factor-wise log losses. Both factors remain supervised, and if $p < q$, the entity factor receives a larger correction because $1-p > 1 - q$. For a negative conjunctive target ($y = 0$),

$$\ell^-_{\text{prod}} = -\log(1 - pq), \qquad \nabla \ell^-_{\text{prod}} = \left(\frac{pq(1 - p)}{1 - pq}, \frac{pq(1 - q)}{1 - pq} \right). \tag{17}$$

Once either factor becomes weak, $pq \to 0$ and both gradients naturally vanish. This agrees with the asymmetric logic of conjunction: one violated condition is already sufficient to invalidate the joint event, so the remaining valid factor need not be further suppressed.

Why Product over Mean and Minimum? For Mean coupling,

$$u_{\text{mean}} = \frac{p + q}{2},$$

the gradients under the same target y are

$$\frac{\partial \ell_{\text{mean}}}{\partial e^I} = \frac{(u_{\text{mean}} - y)p(1 - p)}{2u_{\text{mean}}(1 - u_{\text{mean}})}, \qquad \frac{\partial \ell_{\text{mean}}}{\partial e^E} = \frac{(u_{\text{mean}} - y)q(1 - q)}{2u_{\text{mean}}(1 - u_{\text{mean}})}. \tag{18}$$

Mean is compensatory: a strong response from one factor can maintain a large coupled confidence despite weak support from the other. For a positive target,

$$\nabla \ell^+_{\text{mean}} = \left(-\frac{p(1 - p)}{p + q}, -\frac{q(1 - q)}{p + q} \right). \tag{19}$$

If $q \to 0$, its gradient also approaches zero, although the event factor is precisely the one requiring correction. Conversely, for a negative target,

$$\nabla \ell^-_{\text{mean}} = \left(\frac{p(1 - p)}{2 - p - q}, \frac{q(1 - q)}{2 - p - q} \right), \tag{20}$$

so even when one factor has already become weak enough to invalidate the conjunction, Mean may continue suppressing the other valid factor.

Minimum removes this compensation but introduces a different limitation. Assuming $p < q$, we have $u_{\min} = p$, and away from the non-differentiable tie point,

$$\frac{\partial \ell_{\min}}{\partial e^I} = p - y, \qquad \frac{\partial \ell_{\min}}{\partial e^E} = 0. \tag{21}$$

Minimum therefore behaves as a hard bottleneck: all supervision is routed through the currently weaker factor, while the other factor is discarded from the optimization. Product combines the desirable properties of both alternatives. It is non-compensatory like Minimum, since $pq \leq \min(p, q)$, but remains smooth and jointly responsive to both factors. It therefore acts as a *soft bottleneck*: both branches remain supervised, with stronger correction automatically assigned to the less confident factor, while negative updates vanish once the conjunction has already been invalidated.

A.2 Analysis of Factor-Selective Counterfactual Learning

Conjunctive supervision constrains the joint entity–event evidence but leaves the validity of individual factors underdetermined. FCL addresses this ambiguity by preserving the full 2×2 validity structure and modeling both unary factor evidence and their dependency. We analyze how the structured energy reduces to independent factor modeling when the interaction is removed and how the interaction term improves discrimination among validity configurations.

Independent factor modeling without interaction. Following the notation in the main paper, let A_z, B_z, and C_z denote the temporally pooled entity evidence, event evidence, and same-time interaction for an intervened query q^z, respectively. For a candidate validity state $(r_I, r_E) \in \{-1, +1\}^2$, the structured compatibility score is

$$H_z(r_I, r_E) = \frac{r_I A_z + r_E B_z + \eta r_I r_E C_z}{T_{\text{cf}}}. \tag{22}$$

The unary terms encode entity and event validity independently, whereas the bilinear term captures their dependency. When $\eta = 0$, the interaction term vanishes and the energy reduces to

$$H_z(r_I, r_E) = \frac{r_I A_z + r_E B_z}{T_{\text{cf}}}.$$

Under this additive formulation, the softmax normalization factorizes:

$$\begin{aligned} &\sum_{r_I, r_E} \exp\left(\frac{r_I A_z + r_E B_z}{T_{\text{cf}}}\right) \\ &= \left(\sum_{r_I} \exp \frac{r_I A_z}{T_{\text{cf}}}\right)\left(\sum_{r_E} \exp \frac{r_E B_z}{T_{\text{cf}}}\right). \end{aligned}$$

Consequently, the joint validity distribution becomes

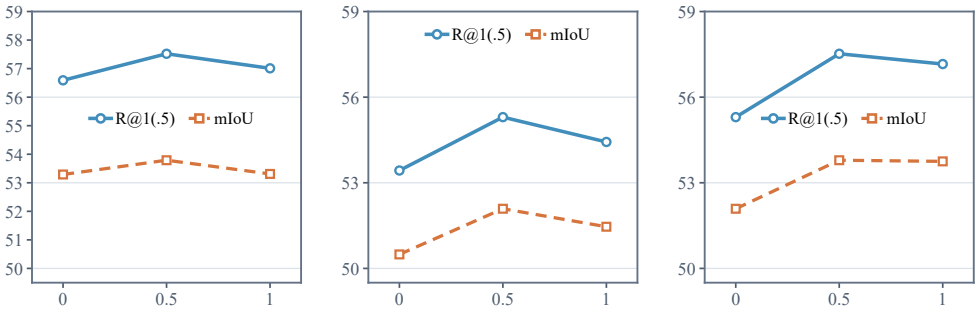
$$P(r_I, r_E \mid q^z) = P_I(r_I \mid q^z) P_E(r_E \mid q^z). \tag{23}$$

Thus, $\eta = 0$ corresponds exactly to independent factor-wise modeling, while the interaction term introduces dependency between entity and event validity.

Interaction-driven state discrimination. To characterize the role of the interaction term, we consider the relative preference between same-validity and mixed-validity states. Let P_{++}, P_{+-}, P_{-+}, and P_{--} denote the probabilities of the four validity configurations. Since all states share the same softmax normalization,

Table 7: Performance–efficiency comparison on IDVTG-Gym. Both methods use identical video and reference-image features pre-extracted offline with CLIP ViT-L/14. The CLIP encoder parameters and offline feature extraction cost are excluded from all reported complexity measurements. All complexity metrics are measured at inference time: GFLOPs are estimated for a single inference forward pass, while latency is measured during inference. FCL is used only as a training-time supervision mechanism and introduces no additional inference branch.

Method	Params (M)	GFLOPs	Latency (ms/video)	mIoU
VGD-Agg†	19.153	10.317	331.03	55.07
Ours (SCG+FCL)	21.265	11.770	351.63	**64.33**
Δ	+2.112	+1.453	+20.60	**+9.26**



(a) Interaction η (b) Synchronization λ_{sync} (c) Counterfactual λ_{cf}

Figure 4: Hyperparameter sensitivity on IDVTG-InternVid. We vary (a) the interaction coefficient η and (c) the counterfactual weight λ_{cf} within the complete SCG+FCL framework, while (b) evaluates the synchronization weight λ_{sync} on SCG-only.

$$\log \frac{P_{++}P_{--}}{P_{+-}P_{-+}} = H_{++} + H_{--} - H_{+-} - H_{-+} = \frac{4\eta C_z}{T_{\text{cf}}}. \tag{24}$$

The unary validity terms cancel completely in this contrast. Hence, the interaction term specifically controls the preference between same-validity states $(++, --)$ and mixed-validity states $(+-, -+)$.

B Sensitivity and Efficiency Analysis

B.1 Computational Overhead

As shown in Table 7, we evaluate the inference-time computational overhead on the same 50 IDVTG-Gym videos with batch size 1. Each configuration is measured three times on a single NVIDIA A800 GPU. Compared with the reproduced VGD-Agg† baseline, our full model improves mIoU by 9.26 points while introducing only 2.112M additional parameters and 1.453 GFLOPs. The measured inference latency increases by 20.60 ms/video. Since FCL is used only during training and introduces no additional inference branch, the test-time overhead comes entirely from the lightweight SCG module. Overall, the results demonstrate that SCG+FCL achieves substantial performance gains with modest inference-time overhead.

Hyperparameter Sensitivity Analysis. We further examine the sensitivity of the three key hyperparameters in Fig. 4. The default setting provides the best overall performance for all three.



Figure 5: Additional examples of factor-selective counterfactual interventions.

For η, the gain over $\eta = 0$ confirms the benefit of structured factor interaction, while nearby settings remain competitive. For λ_{sync}, removing synchronized supervision clearly degrades SCG-only performance, whereas an overly large weight is less effective. For λ_{cf}, $\lambda_{\text{cf}} = 0$ corresponds to SCG-only, and enabling FCL yields a clear improvement with limited sensitivity between 0.5 and 1.0.

C Additional Qualitative Results

C.1 Factor-Selective Counterfactual Visualizations

As shown in Fig. 5, we provide additional examples of the four factor-validity states used by FCL. Each group contains the factual state q^{++} together with the event-intervened q^{+-}, entity-intervened q^{-+}, and dual-intervened q^{--} states. The counterfactuals remain semantically plausible within the same gymnastics domain rather than forming trivial unrelated negatives. Event intervention changes the queried action while preserving the reference entity, whereas entity intervention replaces the reference athlete while keeping the event query and factual ground-truth interval unchanged; the dual intervention combines both changes. This controlled construction makes the constituent factor validity explicit and provides factor-specific supervision beyond the factual conjunction alone.

C.2 Additional Temporal Grounding Visualizations

Figure 6 presents two representative qualitative examples on IDVTG-Gym. The first example shows that SCG improves temporal alignment through synchronized entity–event grounding. The baseline is distracted by an incorrect temporal segment, while SCG shifts the prediction toward the interval where the reference entity and queried event are simultaneously supported. With factor-selective supervision during training, FCL further enhances the learned model's ability to distinguish constituent factor validity, leading to more precise localization. The second example demonstrates a more challenging long-duration grounding scenario. Although the baseline identifies a floor-exercise segment, it is performed by a different gymnast and therefore fails to satisfy the reference-entity condition. SCG effectively suppresses the temporal distractor by calibrating proposal relevance with synchronized entity–event evidence. FCL further improves the learned representation through factor-level supervision, making the model more robust to ambiguous entity–event combinations.



(a) Qualitative example 1.



(b) Qualitative example 2.

Figure 6: Qualitative temporal grounding comparisons on IDVTG-Gym. The reference image specifies the target entity, while the text query describes the target event. Each example compares the temporal predictions of VGD-Agg†, SCG, and SCG+FCL against the ground-truth interval. Frames within the ground-truth interval are highlighted in red.