# GROUNDING AS VERIFICATION: CLOSING THE LOOP IN GROUNDED VIDEO CAPTIONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Grounded video captioning systems generate descriptions with phrase-level localizations, but existing pipelines still flow one way, *caption* $\rightarrow$ *grounding*: grounding reveals where the model points, but never feeds back whether generated entities are visually supported. We argue that grounding should also act as a verifier, forming a *caption* $\leftrightarrow$ *grounding* loop in which localization evidence verifies and refines generation. Current grounding heads, however, typically return only one representative trajectory per phrase. This single-trajectory evidence is too limited for feedback: a plural phrase may look grounded even when the model finds only one instance. We therefore introduce a lightweight *Multi-Instance Adapter* that expands each phrase query into multiple persistent evidence slots, upgrading grounding into a structured multi-instance verifier. We then build annotation-free visual self-alignment with instance-presence and temporal-consistency rewards for KL-regularized policy optimization, improving captioning and grounding without ground-truth captions or extra box supervision during self-alignment. On iGround, the adapter improves AP50 from 40.0 to 51.7 over GROVE, and the full closed-loop model reaches 53.3 AP50 and 88.0 CIDEr. Results on VidSTG and ActivityNet-Entities further show effectiveness beyond the main benchmark. Together, these results support a verification-centered loop in which grounding provides feedback for better captions, and improved captions induce cleaner grounding.

## 1 INTRODUCTION

Video captioning is an important interface for video understanding because it turns complex visual content into natural-language descriptions (Venugopalan et al., 2015; Yao et al., 2015; Xu et al., 2016; Lin et al., 2022; Yang et al., 2023). Yet a free-form caption alone provides limited means to verify *which* visual evidence supports each claim. Caption grounding provides this evidence by localizing the noun phrases or entities mentioned in a given caption, making the description attributable rather than purely textual. However, because the textual input is fixed in advance, caption grounding verifies an existing description rather than deciding which entities or events should be described.

Image-based grounded captioning has begun to close this gap by producing captions together with phrase-aligned regions (Rasheed et al., 2024; Zhang et al., 2023; Peng et al., 2023; Ma et al., 2024; Patel et al., 2022). The video counterpart requires both generating a video-level caption and predicting temporally dense, consistent boxes for multiple noun phrases in that caption (Zanfir et al., 2016; Ma et al., 2020; Zhou et al., 2019). This is substantially harder than the image setting because objects may move, disappear under occlusion, and reappear across frames. As a result, most video grounding work has remained query-driven: given a caption or query, the model localizes the described target over time rather than generating and grounding the caption together (Zhang et al., 2020; Tang et al., 2022; Yang et al., 2022).

More recently, GROVE (Kazakos et al., 2025) has made an important step toward this joint video setting. With large-scale phrase-grounding pretraining, it generates video-level captions while grounding noun phrases with temporally dense boxes across frames, thereby exposing localization evidence for each generated entity. Yet this evidence is still used only as a terminal prediction: caption hidden states are converted into grounding queries and decoded into boxes, so information flows one way, *caption* $\rightarrow$ *grounding*, with no return path. Weak, incomplete, or temporally inconsistent



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.

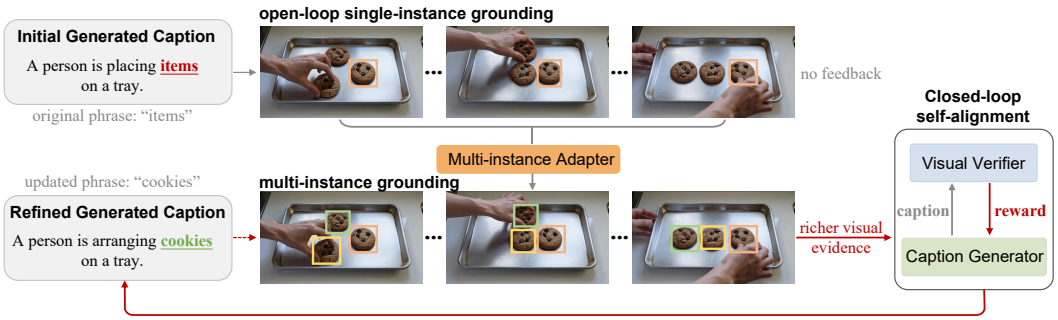


Figure 1: Closed-loop overview. **Top:** Open-loop systems bind each phrase to one grounded instance; the baseline says *"items"*, tracks one cookie, and receives no grounding feedback. **Bottom:** Our Multi-Instance Adapter expands grounding into multiple trajectories, supporting the refined phrase *"cookies"* and cleaner grounding. **Right:** A visual verifier scores sampled captions by presence and temporal consistency, completing the *caption* $\leftrightarrow$ *grounding* loop.

grounding is a strong cue that a generated phrase lacks visual support, but current systems do not feed it back to verify or refine generation.

Realizing this *caption* $\leftrightarrow$ *grounding* feedback loop requires the grounding signal to be both rich enough to serve as reliable evidence and convertible into an effective learning signal for the discrete caption generator. Current grounding heads fall short on both counts. First, they produce a single representative hypothesis per phrase, collapsing multiple relevant instances into one box trajectory. A single representative box cannot tell whether a phrase such as *"cookies"* is supported by several relevant objects or only by one convenient instance, leaving the verifier unable to distinguish correct grounding from under-grounding. As illustrated in Figure 1, the open-loop baseline tracks only one cookie, discarding the multi-instance evidence needed to judge whether the phrase is well-supported. To serve as verification evidence, the grounding interface must expose multiple instance trajectories per phrase so that the breadth and temporal consistency of visual support can be assessed. Second, even with richer evidence, caption generation is discrete and non-differentiable, so grounding scores cannot be directly backpropagated to the generated words. The challenge is therefore to convert structured visual evidence into a stable, non-differentiable yet effective reward signal that can guide caption refinement.

We address both challenges in a unified framework built on GROVE (Kazakos et al., 2025), whose large-scale pretrained phrase-conditioned grounding interface provides a strong starting point. To build a verification-ready grounding signal, we insert a lightweight *Multi-Instance Adapter* that expands each phrase query into multiple persistent evidence slots through anchored residual slot embeddings trained with trajectory-level Hungarian set supervision. This slot-structured upgrade preserves the original phrase-query geometry while exposing multiplicity and temporal support, turning the grounding branch from a single existential prediction into a structured set of candidate evidence tracks, as illustrated in Figure 1. To close the loop, we then keep a frozen reference copy of the upgraded verifier stack and reuse it as a task-native visual critic. As sketched on the right of Figure 1, sampled captions are replayed through this frozen verifier to produce a visual reward based on multi-instance presence and temporal consistency; the reward is fed back to the generator through policy optimization with KL regularization, refining an initial coarse description such as *"placing items"* into a more specific and better-supported caption such as *"arranging cookies"*. As the red feedback arrow in Figure 1 highlights, this refinement also leads to cleaner caption-conditioned grounding, completing the *caption* $\leftrightarrow$ *grounding* loop between better descriptions and better verification, without ground-truth captions or additional box supervision.

We validate the proposed framework on three grounding benchmarks: iGround (Kazakos et al., 2025), VidSTG (Zhang et al., 2020), and ActivityNet-Entities (Zhou et al., 2019). Across all three, we observe consistent grounding improvements over strong GROVE-based baselines, while caption gains and detailed self-alignment analysis are reported on iGround. Together, these results suggest that richer grounding evidence and closed-loop self-alignment are complementary: multi-instance evidence delivers the main grounding gain, while caption-side refinement additionally improves both faithfulness and phrase-grounding quality. Our contributions are:

1. We reformulate grounded video captioning as a *caption* $\leftrightarrow$ *grounding* closed loop, where phrase-level localization feeds back to verify and refine generation instead of remaining a terminal prediction.
2. We instantiate this loop with a lightweight *Multi-Instance Adapter* that upgrades a pretrained single-instance grounding interface into a structured multi-instance verifier through anchored residual slot embeddings and trajectory-level set supervision, then reuse the verifier as an annotation-free visual critic for policy-gradient caption refinement.
3. We validate the framework on iGround, where the adapter raises AP50 from 40.0 to 51.7 and closed-loop self-alignment further reaches 53.3 AP50 and 88.0 CIDEr while reducing object hallucination from 22.8% to 18.7%; gains transfer consistently to VidSTG and ActivityNet-Entities, supporting generalization beyond the main benchmark.

## 2 Related Work

**Image-based grounded captioning.** Recent multimodal models have advanced grounded caption generation in still images toward phrase-level localization, with approaches ranging from explicit location tokens or boxes (Peng et al., 2023; Chen et al., 2023; Zhang et al., 2023), to external grounding modules (Zhao et al., 2023), to tighter integration of dense grounding into the captioning stack (Rasheed et al., 2024; Ma et al., 2024). These works establish language hidden states as effective interfaces for visual attribution, but they operate on static images and do not address temporal consistency or video-level caption generation.

**Spatio-temporal grounding in video.** One line of work studies grounded video description and captioning under varying levels of supervision (Zanfir et al., 2016; Ma et al., 2020; Zhou et al., 2019; 2023). A larger body of work addresses spatio-temporal grounding of a given text query, including supervised transformer-based methods (Zhang et al., 2020; Tang et al., 2022; Yang et al., 2022; Su et al., 2021; Jin et al., 2022; Lin et al., 2023; Gu et al., 2024), self-supervised approaches for instructional video (Tan et al., 2021; Chen et al., 2024), and open-vocabulary grounded video models (Wasim et al., 2024; Munasinghe et al., 2023; 2025). These systems are primarily query-driven: they localize a pre-existing query and typically return a single target tube, rather than jointly generating captions and grounding them. Among benchmarks, ActivityNet-Entities (Zhou et al., 2019) annotates noun phrases in captions but grounds each object on only a single frame, while VidSTG and HC-STVG (Zhang et al., 2020; Tang et al., 2022) provide temporally dense grounding but assume the query text is given. GROVE (Kazakos et al., 2025) is the closest baseline to our setting, combining phrase-conditioned captioning and grounding with large-scale pretraining on HowToGround1M. We build on GROVE's pretrained interface but upgrade it from a single-instance prediction head to a multi-instance evidence mechanism that can also serve as a verifier.

**Sequence-level training and alignment for faithful generation.** Reinforcement learning has long been used in caption generation to optimize sequence-level objectives beyond teacher forcing, most notably through self-critical sequence training (Rennie et al., 2017) and reward-based video captioning (Pasunuru & Bansal, 2017). A related line addresses hallucination and visual faithfulness in image captioning (Biten et al., 2022; Dai et al., 2023), in large vision-language models (Li et al., 2023; Lee et al., 2024), and through multimodal alignment with human or AI feedback (Yu et al., 2024; 2025); benchmarks such as VidHalluc further reveal temporal hallucination failures specific to video (Li et al., 2025). These works highlight the need for feedback beyond token likelihood, but existing approaches typically rely on text-only matching metrics, external preference signals, or post-hoc language reasoning. Our approach instead reuses the model's own upgraded grounding branch as a task-native spatio-temporal verifier and stabilizes optimization with a frozen reference policy, so that the self-alignment stage is guided by localized visual evidence without requiring ground-truth captions or additional box supervision.

## 3 Method

### 3.1 Framework Overview

As shown in Figure 2, our framework starts from the GROVE checkpoint pretrained on HowToGround1M (Kazakos et al., 2025). GROVE marks each noun phrase with a `[DET]` token, projects

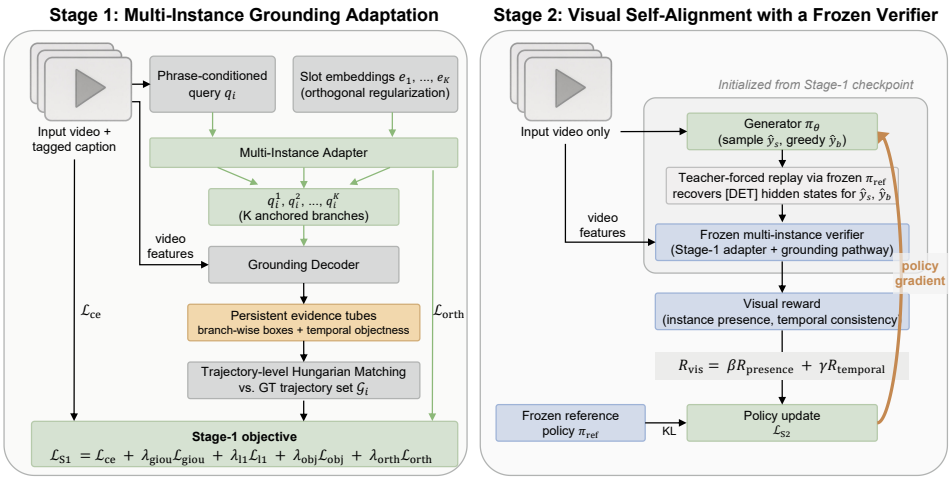


Figure 2: Method overview. **Stage 1** expands each phrase-conditioned query into $K$ anchored slots and trains persistent evidence tubes with caption, grounding, objectness, and orthogonality losses. **Stage 2** uses video only: sampled captions are replayed through a frozen reference verifier to compute presence and temporal rewards, and a KL-anchored policy update refines the generator.

the corresponding hidden state into a phrase-conditioned grounding query, and decodes it with video features for localization; Section 3.2 reviews this interface. We then upgrade the model in two stages. In the first stage, we add a lightweight *Multi-Instance Adapter* that expands each phrase-conditioned query into $K$ anchored slots and trains the resulting evidence tubes with trajectory-level set matching and orthogonality regularization (Section 3.3). In the second stage, we initialize the caption policy from the Stage-1 checkpoint and use a frozen reference copy to replay sampled and greedy captions, recover their `[DET]` states, score them with the frozen multi-instance verifier, and provide both visual rewards and a KL anchor for policy optimization (Section 3.4). Section 3.5 gives the optimization and implementation details.

### 3.2 Preliminaries: Phrase-Conditioned Grounding

Given a video and a tagged caption containing $N$ noun phrases, where each noun phrase is followed by a special `[DET]` token, let $h_i \in \mathbb{R}^{d_h}$ be the hidden state of the $i$-th `[DET]` token. GROVE projects it into the grounding query $q_i = W_{\text{proj}} h_i + b_{\text{proj}} \in \mathbb{R}^{d_q}$, which is fed into the grounding decoder to predict, at each frame $t$, one bounding box together with one temporal objectness score $(\hat{b}_{i,t}, \hat{o}_{i,t})$ for phrase $i$.

This single-instance representation is sufficient for phrase-level existence prediction, but it is too narrow for verification: it can trace one representative support tube, but cannot tell whether all visually relevant instances of a phrase are covered. Therefore, the rest of the method upgrades this interface from one hypothesis per phrase to multiple hypotheses per phrase.

### 3.3 Stage 1: Multi-Instance Grounding Adaptation

To expose multi-instance evidence while retaining the pretrained phrase-query geometry, we introduce a lightweight Multi-Instance Adapter. For each phrase-conditioned query $q_i$, we create $K$ slot queries $q_i^k = q_i + e_k$, where $e_k \in \mathbb{R}^{d_q}$ is a learnable slot embedding shared across phrases. This anchored residual parameterization keeps $q_i$ as the semantic anchor while the shared offsets $\{e_k\}_{k=1}^K$ create reusable query-space directions for different instance hypotheses. We orthogonally initialize the slot embeddings and keep them separated during fine-tuning with

$$\mathcal{L}_{\text{orth}} = \left\| \bar{E}\bar{E}^\top - I_K \right\|_F^2 . \tag{1}$$

Here, $\bar{E} \in \mathbb{R}^{K \times d_q}$ stacks the normalized embeddings $\bar{e}_k = e_k / \|e_k\|_2$ as rows.

The expanded slot queries are then passed through the existing GROVE grounding pathway. At frame $t$, slot $k$ predicts a bounding box $\hat{b}_{i,t}^k$ and temporal objectness logit $\hat{o}_{i,t}^k$; reusing the same slot index

across sampled frames gives the predicted slot tube $\hat{\tau}_i^k = \{(\hat{b}_{i,t}^k, \hat{o}_{i,t}^k)\}_{t=1}^T$, one candidate instance trajectory for phrase $i$.

To make each slot temporally coherent, we do not match predictions independently at each frame. Instead, for phrase $i$ we collect ground-truth trajectories $\mathcal{G}_i = \{g_i^1, \ldots, g_i^{J_i}\}$, where $J_i$ is the number of distinct annotated trajectories in the sampled clip. For trajectory $j$, let $v_{i,t}^j \in \{0,1\}$ indicate visibility in frame $t$, $\mathcal{V}_i^j = \{t \mid v_{i,t}^j = 1\}$ be its visible-frame set, and $g_i^j = \{b_{i,t}^j\}_{t \in \mathcal{V}_i^j}$. We then compare each predicted slot tube against each ground-truth trajectory as a whole and perform trajectory-level Hungarian matching (Kuhn, 1955) between the $K$ slot tubes and the $J_i$ ground-truth trajectories using the clip-level cost

$$C_{kj}^{\text{traj}} = \frac{1}{|\mathcal{V}_i^j|} \sum_{t \in \mathcal{V}_i^j} \Big(1 - \text{IoU}(\hat{b}_{i,t}^k, b_{i,t}^j)\Big). \tag{2}$$

The cost averages the per-frame IoU disagreement between slot $k$ and trajectory $j$ over visible frames, so assignment is decided at the clip level rather than frame by frame. The Hungarian solver produces a one-to-one partial matching $\mathcal{M}_i \subseteq \{1, \ldots, K\} \times \{1, \ldots, J_i\}$. If $(k, j) \in \mathcal{M}_i$, slot $k$ is assigned to trajectory $j$ for the entire sampled clip; its slot-level visibility target is $y_{i,t}^k = v_{i,t}^j$, and otherwise $y_{i,t}^k = 0$. For a matched slot, frames outside its assigned visible-frame set receive $y_{i,t}^k = 0$, so spurious activations there are suppressed by the temporal objectness branch. If $J_i > K$, at most $K$ trajectories can be assigned, and the remaining trajectories fall outside the fixed slot budget for that sampled clip. Matched slots receive box regression supervision through $\mathcal{L}_{\text{giou}}$ and $\mathcal{L}_{\text{l1}}$ on the frames where $y_{i,t}^k = 1$, while unmatched slots are treated as negatives by the temporal objectness branch. For a training sample with $N$ tagged phrases, $T$ sampled frames, and $K$ grounding slots per phrase, the temporal objectness loss is

$$\mathcal{L}_{\text{obj}} = \frac{1}{NTK} \sum_{i,t,k} \text{BCE}\big(\hat{o}_{i,t}^k, y_{i,t}^k\big)\,. \tag{3}$$

The complete Stage 1 objective is

$$\mathcal{L}_{\text{S1}} = \mathcal{L}_{\text{ce}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{l1}}\mathcal{L}_{\text{l1}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}}, \tag{4}$$

where $\mathcal{L}_{\text{ce}}$ is the teacher-forced cross-entropy on tagged captions, and $\mathcal{L}_{\text{giou}}$ and $\mathcal{L}_{\text{l1}}$ follow the standard DETR-style GIoU+L1 box regression formulation (Carion et al., 2020; Rezatofighi et al., 2019), averaged over matched $(k, t)$ pairs with $y_{i,t}^k = 1$. Together, the shared slot embeddings, trajectory-level Hungarian supervision, and orthogonality regularization bias the $K$ slots toward distinct instance hypotheses rather than near-duplicate solutions. Because each matched slot is tied to a unique ground-truth trajectory over the sampled clip, the resulting slot tubes form time-consistent evidence tracks.

This multi-instance interface changes grounding from a single existential prediction into a structured evidence set. Stage 2 uses this set along two complementary axes: instance presence, a soft support measure over active slots, and temporal consistency, which measures whether each active slot remains stable across frames.

### 3.4 Stage 2: Visual Self-Alignment with a Frozen Verifier

Once Stage 1 has produced structured multi-instance evidence, we keep a frozen reference copy $\pi_{\text{ref}}$ of the Stage-1 model and reuse its visual encoder and grounding modules downstream of the `[DET]` hidden states as a frozen verifier for the trainable caption policy $\pi_\theta$. Unlike RL captioning methods that optimize external text metrics such as CIDEr, our reward is task-native and visually grounded. Using the same frozen model for replay-based verification and for KL anchoring keeps the visual reward and the language prior tied to the same supervised reference.

For a video $x$, the policy $\pi_\theta$ produces a sampled caption $\hat{y}_s$ and a greedy baseline $\hat{y}_b$. The policy itself remains trainable in Stage 2, so its induced `[DET]` hidden states may change during optimization. For reward computation, however, the policy generates captions in the same tagged format learned in Stage 1, and replaying either sequence through $\pi_{\text{ref}}$ in teacher-forcing mode recovers reward-side `[DET]` hidden states for the generated noun phrases. We score those states with the frozen verifier

using $R_{\text{vis}} = \beta R_{\text{presence}} + \gamma R_{\text{temporal}}$, where $\beta$ and $\gamma$ balance instance-presence evidence and local temporal persistence. We consider only captions with at least one recoverable tagged noun phrase; samples with $N = 0$ are skipped. Using the frozen reference model for reward evaluation keeps the reward path fixed during policy optimization. At inference time, this replay path is not used: the current policy generates the tagged caption, and its own `[DET]` hidden states are passed to the frozen Stage-1 grounding heads to produce localizations.

**Instance presence.** We define

$$R_{\text{presence}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \sigma(\hat{o}_{i,t}^{k}), \tag{5}$$

where $N$ is the number of recovered noun phrases, $T$ is the number of sampled frames, and $\sigma(\cdot)$ is the sigmoid function. This term averages verifier objectness over phrases and frames while summing over the $K$ evidence slots, so $R_{\text{presence}} \in [0, K]$ serves as a soft multi-instance support signal rather than an exact cardinality estimate. We keep the sum over slots so that multiple visually supported slots can increase phrase support; the coefficient $\beta$ calibrates its scale relative to the temporal term. The normalization over noun phrases also reduces the incentive to add unsupported entities, with the KL anchor below providing an additional language-side constraint.

**Temporal consistency.** For phrase $i$ and slot $k$, recall the predicted slot tube $\hat{\tau}_i^k = \{(\hat{b}_{i,t}^k, \hat{o}_{i,t}^k)\}_{t=1}^T$ from Stage 1. We define its binary visibility mask as $m_{i,t}^k = \mathbb{I}\big[\sigma(\hat{o}_{i,t}^k) > \tau_{\text{vis}}\big]$, where $\tau_{\text{vis}}$ is the visibility threshold. We denote the active slots for phrase $i$ by $\mathcal{S}_i = \Big\{k \mid \sum_{t=1}^T m_{i,t}^k > 0\Big\}$ and the active adjacent-frame transitions of tube $\hat{\tau}_i^k$ by $\mathcal{A}_i^k = \big\{t \mid 1 \le t < T,\ m_{i,t}^k + m_{i,t+1}^k > 0\big\}$. Intuitively, $\mathcal{C}(\hat{\tau}_i^k)$ asks whether an active slot remains both visible and geometrically stable from one sampled frame to the next: transitions visible in both frames contribute their box IoU, while one-sided visible transitions contribute zero. The tube consistency is

$$\mathcal{C}(\hat{\tau}_i^k) = \frac{1}{\max(1, |\mathcal{A}_i^k|)} \sum_{t \in \mathcal{A}_i^k} \mathbb{I}\big[m_{i,t}^k = m_{i,t+1}^k = 1\big] \cdot \text{IoU}(\hat{b}_{i,t}^k, \hat{b}_{i,t+1}^k). \tag{6}$$

Here, $\mathcal{C}(\hat{\tau}_i^k)$ is the average adjacent-frame self-IoU of slot $k$ over the transitions where it is active. If the slot disappears in one of two adjacent frames, that transition contributes zero but is still included in the average, so unstable tracks receive lower scores. We set $\mathcal{C}(\hat{\tau}_i^k) = 0$ when $\mathcal{A}_i^k = \emptyset$. The temporal reward then averages this slot-level persistence over phrases and over the active slots of each phrase,

$$R_{\text{temporal}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\max(1, |\mathcal{S}_i|)} \sum_{k \in \mathcal{S}_i} \mathcal{C}(\hat{\tau}_i^k). \tag{7}$$

If no slot is active for phrase $i$, the inner sum is empty and that phrase contributes zero. This normalization prevents captions from receiving larger temporal reward merely by activating more slots. Because Stage 1 trains each slot to track a single trajectory identity over the sampled clip, slot-wise self-IoU is meaningful rather than an artifact of anonymous per-frame detections.

**Policy optimization with KL regularization.** Optimizing only for $R_{\text{vis}}$ would encourage reward hacking and drift from the supervised language prior. Following reference-model KL regularization in language-policy alignment (Ziegler et al., 2019), we use the same frozen reference model $\pi_{\text{ref}}$ as a length-normalized KL anchor:

$$\mathcal{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}; \hat{y}) = \frac{1}{L} \sum_{\ell=1}^{L} \text{KL}(\pi_\theta(\cdot \mid x, \hat{y}_{<\ell}) \,\|\, \pi_{\text{ref}}(\cdot \mid x, \hat{y}_{<\ell})), \tag{8}$$

where $L$ is the generated length and $\hat{y}_{<\ell}$ is the prefix before token $\ell$. This token-level KL term measures how far the trainable policy's next-token distribution moves from the frozen reference under the same prefix, discouraging the captioner from drifting away from the supervised model while it optimizes the visual reward. Unlike the visual reward, this term is fully differentiable and is optimized directly.

We optimize the generator with a self-critical REINFORCE estimator (Rennie et al., 2017), using the greedy caption $\hat{y}_b$ as a low-variance baseline for the sampled caption $\hat{y}_s$. Both reward evaluations are detached from gradient flow, and we use the clipped stop-gradient advantage $A_{\text{vis}} = \text{clip}(R_{\text{vis}}(\hat{y}_s) - R_{\text{vis}}(\hat{y}_b), -\delta, \delta)$, where $\delta > 0$ limits high-variance reward updates. Using the autoregressive factorization $\log \pi_\theta(\hat{y}_s|x) = \sum_{\ell=1}^{L} \log \pi_\theta(\hat{y}_{s,\ell} \mid x, \hat{y}_{s,<\ell})$, the Stage-2 loss is

$$\mathcal{L}_{\text{S2}} = -\frac{A_{\text{vis}}}{L} \log \pi_\theta(\hat{y}_s|x) + \lambda_{\text{KL}} \mathcal{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}; \hat{y}_s). \tag{9}$$

We use REINFORCE because the verifier supplies a single sequence-level non-differentiable reward, while the greedy caption already provides a natural self-critical baseline and $\pi_{\text{ref}}$ supplies an explicit KL anchor. The length-normalized log-probability keeps the REINFORCE term comparable across captions of different lengths and numerically compatible with the KL term. Together, the intrinsic visual critic and the frozen reference policy enable Stage 2 to strengthen multi-instance visual support, suppress unsupported entities, and encourage temporally stable support without any text-side reward.

### 3.5 Implementation Details

Stage 1 initializes from the GROVE checkpoint pretrained on HowToGround1M (Kazakos et al., 2025), and Stage 2 initializes from the resulting Stage-1 checkpoint. Following GROVE, the base architecture builds on GLaMM (Rasheed et al., 2024) with CLIP-L video features (Radford et al., 2021), SAM-based localization modules (Kirillov et al., 2023), and a Vicuna-7B language model (Chiang et al., 2023). Unless stated otherwise, we use $K = 3$ slots, sample 8 frames at image size 512, and train on two NVIDIA A100 80GB GPUs. Stage 1 performs supervised multi-instance adaptation, while Stage 2 freezes the verifier and updates only the language-side generator with the visual reward and KL anchor. Full optimizer settings, loss weights, trainable-parameter choices, decoding settings, and reward hyperparameters are reported in Appendix A.

## 4 Experiments

### 4.1 Datasets and Metrics

We evaluate on iGround (Kazakos et al., 2025), VidSTG (Zhang et al., 2020), and ActivityNet-Entities (Zhou et al., 2019), following the same dataset-specific protocols used in GROVE.

**iGround.** Following GROVE, we report METEOR (Banerjee & Lavie, 2005) and CIDEr (Vedantam et al., 2015) for caption quality, and AP50 and Recall for grounding quality. Recall counts a prediction as positive only when both noun-phrase similarity and box IoU pass the required thresholds; AP50 and Recall are computed per video and averaged across videos. Unless stated otherwise, iGround numbers use all-frame evaluation. For diagnostics, we report AP50 on singular phrases (Sing.-cond.) and non-singular phrases (Non-sing.-cond.), and object hallucination (Obj. Halluc.) as the percentage of generated noun phrases that fail to semantically match any annotated entity using Sentence-BERT `all-MiniLM-L6-v2` at cosine threshold 0.7, analogous to instance-level CHAIR (Rohrbach et al., 2018).

**ActivityNet-Entities.** Following (Zhou et al., 2019), we report $\text{F1}_{\text{all}}$, $\text{F1}_{\text{all_per_sent}}$, $\text{F1}_{\text{loc}}$, and $\text{F1}_{\text{loc_per_sent}}$. $\text{F1}_{\text{all}}$ requires both exact noun-phrase prediction and correct localization with IoU $>$ 0.5, whereas $\text{F1}_{\text{loc}}$ evaluates localization alone. The "per_sent" variants average over sentences rather than noun categories. Evaluation is conducted on the annotated frames provided by the benchmark.

**VidSTG.** Following (Su et al., 2021; Yang et al., 2022; Jin et al., 2022; Zhou et al., 2023), we assume ground-truth temporal localization and report $\text{m}_{\text{sIoU}}$, the average IoU between predicted and ground-truth boxes across frames in the target segment. This protocol evaluates single-target query-conditioned grounding inside the annotated temporal window. We report both declarative and interrogative sentence settings, consistent with the GROVE evaluation.

Detailed split usage, benchmark-specific training and inference protocols, and computational overhead are provided in Appendix B.

### 4.2 Main Results on iGround

Table 1 shows that the supervised Multi-Instance Base delivers the main grounding gain while preserving caption metrics, so the adapter strengthens the evidence interface without destabilizing generation. The larger gain in AP50 than Recall is expected: Recall mainly asks whether a phrase has at least one acceptable match, while AP50 is more sensitive to cleaner boxes and better separation among multiple instances. Closed-loop self-alignment then improves both grounding and captioning, with caption-side gains reflected in CIDEr and the hallucination diagnostic in Table 2. We compare directly with GROVE-family systems here because they share the end-to-end setting; fixed-query video grounding models are reported separately.

Table 1: Main iGround results under the all-frame evaluation protocol.

| | | Grounding | | Captioning | |
|---|---|---|---|---|---|
| Split | Method | AP50 ↑ | Recall ↑ | METEOR ↑ | CIDEr ↑ |
| Test | GROVE baseline (Kazakos et al., 2025) | 40.0 | 28.7 | 21.4 | 83.5 |
| | Ours (Multi-Instance Base) | 51.7 | 30.1 | 21.5 | 83.6 |
| | **Ours (Full Model)** | **53.3** | **30.5** | **22.3** | **88.0** |
| Validation | GROVE baseline (Kazakos et al., 2025) | 42.0 | 26.9 | 19.7 | 92.6 |
| | Ours (Multi-Instance Base) | 52.7 | 28.4 | 19.9 | 93.8 |
| | **Ours (Full Model)** | **53.9** | **29.7** | **20.5** | **97.1** |

Table 2 checks entity-level behavior. The multi-instance base helps most on non-singular phrases, where a single-trajectory head often covers only one object. Closed-loop self-alignment further reduces hallucination and improves both singular and non-singular AP50, showing that entity mentions become better visually supported rather than merely scoring higher under caption metrics.

Table 2: Fine-grained iGround validation diagnostics. Sing.-cond./Non-sing.-cond. report AP50 on singular/non-singular phrases; Obj. Halluc. reports object hallucination.

| Model | Sing.-cond. ↑ | Non-sing.-cond. ↑ | Obj. Halluc. ↓ |
|---|---|---|---|
| GROVE baseline | 45.8 | 16.3 | 22.8 |
| Ours (Multi-Instance Base) | 48.2 | 35.3 | 21.4 |
| Ours (Full Model) | **49.2** | **36.5** | **18.7** |

### 4.3 Results on Other Grounding Benchmarks

Table 3 tests transfer beyond iGround. VidSTG fixes the query sentence and expects one output tube, so its gains show that multi-instance training improves the grounding interface even without caption refinement. ActivityNet-Entities provides only sparse frame annotations, yet all F1 metrics improve, suggesting that the learned instance-support signal remains useful when dense temporal consistency is unavailable.

Table 3: Results on other grounding benchmarks. VidSTG uses given-query dense tube grounding; ActivityNet-Entities uses sparse frame-level phrase grounding.

| VidSTG Declarative | | VidSTG Interrogative | | ActivityNet-Entities | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | $m_{sIoU}$ | Method | $m_{sIoU}$ | Method | $F1_{all}$ | $F1_{sent}$ | $F1_{loc}$ | $F1_{loc/sent}$ |
| STVGBert (Su et al., 2021) | 47.3 | PG-V-L (13B) (Munasinghe et al., 2023) | 35.1 | GVD (Zhou et al., 2019) | 7.1 | 17.3 | 23.8 | 59.2 |
| TubeDETR (Yang et al., 2022) | 59.0 | GLaMM (Rasheed et al., 2024) | 38.6 | GROVE (Kazakos et al., 2025) | 13.4 | 24.1 | 45.0 | 77.3 |
| STCAT (Jin et al., 2022) | 61.7 | VideoGLaMM (Munasinghe et al., 2025) | 39.7 | Ours | **16.6** | **27.8** | **50.4** | **80.6** |
| DenseVOC (Zhou et al., 2023) | 61.9 | GROVE (Kazakos et al., 2025) | 55.5 | | | | | |
| GROVE (Kazakos et al., 2025) | 63.7 | Ours | **57.9** | | | | | |
| Ours | **65.2** | | | | | | | |

### 4.4 Ablation Study

**Stage-1 adapter ablations.** The top block of Table 4 shows that the gain is not from query capacity alone: the single-slot adapter retains the one-instance bottleneck, while naively duplicated queries

add capacity without specialization. Trajectory-level matching outperforms frame-level matching because it preserves slot identity across frames, and orthogonality reduces duplicate evidence tracks.

**Stage-2 self-alignment ablations.** The bottom block shows that CIDEr-based SCST improves text scores but weakens grounding. Our visual reward gives the best joint tradeoff because it combines instance-presence and temporal-consistency evidence; presence-only and temporal-only variants are weaker, the KL anchor limits language drift, and single-slot scoring is inferior. Appendix C supports the default orthogonal initialization, $K = 3$, and $\lambda_{\mathrm{KL}} = 0.20$.

Table 4: Core iGround validation ablations for the Stage-1 adapter and Stage-2 self-alignment.

| | | Grounding | | Captioning | |
|---|---|---|---|---|---|
| Module | Variant | AP50 ↑ | Recall ↑ | METEOR ↑ | CIDEr ↑ |
| Multi-Instance Base | single-slot adapter ($K = 1$) | 43.5 | 27.2 | 19.8 | 92.9 |
| | naive duplicated queries | 40.6 | 25.7 | 19.3 | 91.4 |
| | frame-level Hungarian matching | 50.2 | 27.7 | 19.8 | 93.4 |
| | w/o orthogonality loss | 48.9 | 27.1 | 19.8 | 93.3 |
| | full adapter | **52.7** | **28.4** | **19.9** | **93.8** |
| Visual Self-Alignment | standard SCST (CIDEr reward) | 52.1 | 28.0 | 20.2 | 96.0 |
| | no KL anchor | 51.5 | 27.7 | 18.7 | 91.4 |
| | w/o visual advantage clip | 53.5 | 29.3 | 20.3 | 96.6 |
| | presence-only visual reward | 53.4 | 29.2 | 20.3 | 96.4 |
| | temporal-only visual reward | 53.0 | 28.8 | 20.1 | 95.6 |
| | single-slot reward | 52.9 | 28.5 | 20.0 | 94.4 |
| | full visual self-alignment | **53.9** | **29.7** | **20.5** | **97.1** |

### 4.5 Qualitative Results

Figure 3 shows action correction, hallucination suppression, plural-noun refinement, and finer grounding, indicating visual support rather than mere paraphrasing.

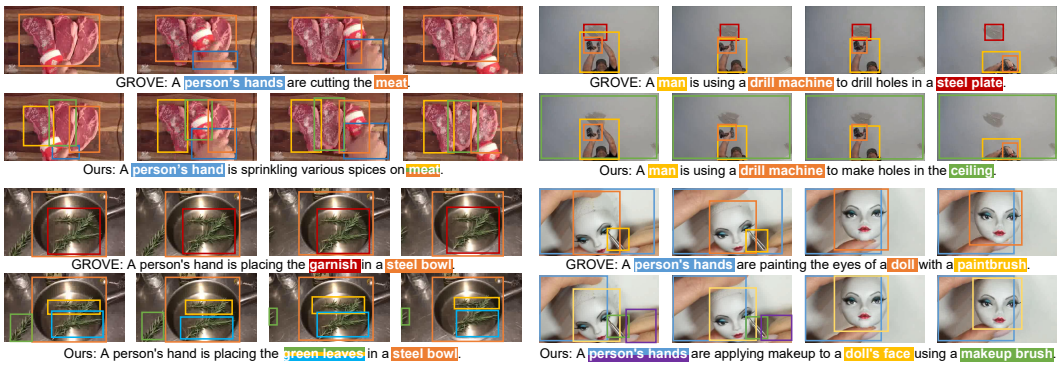


Figure 3: Qualitative iGround comparisons showing action correction, hallucination suppression, plural-noun refinement, and fine-grained grounding. Text colors match box colors; mixed colors denote phrases grounded by multiple trajectories.

## 5 Conclusion

In this work, we have proposed a closed-loop framework for grounded video captioning. Based on the pretrained GROVE interface, we first introduce a lightweight Multi-Instance Adapter that expands each phrase query into multiple evidence slots, allowing the grounding branch to represent richer instance-level support than a single representative trajectory. We then reuse the upgraded verifier as a frozen visual critic, whose instance-presence and temporal-consistency rewards guide KL-regularized self-alignment without ground-truth captions or additional box supervision. Experiments on iGround show that the adapter provides the main grounding gain and that the full model improves both grounding and caption quality while reducing hallucination. Results on VidSTG and ActivityNet-Entities further verify that the learned evidence interface generalizes beyond the main benchmark. Through this work, we hope to encourage grounding to serve not only as an output of generation, but also as feedback for improving generation itself.

## AI USE STATEMENT

Generative AI tools were used to assist with language editing and improve the readability of the manuscript. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

## REFERENCES


Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1381–1390, January 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Daniel Kondermann, Samuel Thomas, Shih-Fu Chang, Rogerio Feris, James Glass, and Hilde Kuehne. What when and where? self-supervised spatio-temporal grounding in untrimmed multi-action videos from narrated instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18419–18429, June 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2136–2148, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.156.

Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Evangelos Kazakos, Cordelia Schmid, and Josef Sivic. Large-scale pre-training for grounded video caption generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24434–24444, October 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 391–404, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.23.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13723–13733, June 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. SwinBERT: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17949–17958, June 2022.

Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. Learning to generate grounded visual captions without localization supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 353–370, 2020.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024.

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm : A large multimodal model for pixel-level visual grounding in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19036–19046, June 2025.

Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 979–985, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1103.

Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, June 2024.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7008–7024, July 2017.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Rui Su, Qian Yu, and Dong Xu. STVGBert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Reuben Tan, Bryan A. Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2022. doi: 10.1109/TCSVT.2021.3085907.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence: Video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4534–4542, December 2015.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, June 2016.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16442–16453, June 2022.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10714–10726, June 2023.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4507–4515, December 2015.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, June 2024.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995, June 2025.

Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2016.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10668–10677, June 2020.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. BuboGPT: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6578–6587, June 2019.

X. Zhou, A. Arnab, C. Sun, and C. Schmid. Dense video object captioning from disjoint supervision. *arXiv preprint arXiv:2306.11729*, 2023.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

# A Training and Implementation Details

**Architecture and shared settings.** Stage 1 initializes from the GROVE checkpoint pretrained on HowToGround1M (Kazakos et al., 2025), and Stage 2 initializes from the resulting Stage-1 checkpoint. Following GROVE, the base architecture builds on GLaMM (Rasheed et al., 2024) and combines a CLIP-L (Radford et al., 2021) global video encoder for captioning, a SAM-based grounding video encoder for fine-grained localization, a SAM-derived box decoder (Kirillov et al., 2023), and a Vicuna-7B language model (Chiang et al., 2023). GROVE further inserts spatio-temporal adapters into the visual encoders and uses a per-frame box head together with a temporal objectness head for visibility prediction. Unless stated otherwise, the one-to-many grounding interface uses $K = 3$ slots. Training and validation use 8 frames, image size 512, and `model_max_length`=1536 on two NVIDIA A100 80GB GPUs.

**Stage 1 optimization.** The supervised multi-instance base is trained for 5 epochs with learning rate $3 \times 10^{-5}$, 100 warmup steps, caption-loss weight 1.0, `train_mask_decoder` enabled, and LoRA disabled. The loss weights are $\lambda_{\text{giou}} = 2.0$, $\lambda_{\text{l1}} = 2.0$, $\lambda_{\text{obj}} = 2.0$, and $\lambda_{\text{orth}} = 0.5$. Following GROVE's fine-tuning protocol, Stage 1 optimizes the CLIP/SAM adapters, the SAM mask decoder, the grounding heads, the text-to-grounding projection, the multimodal projector, `embed_tokens`, and `lm_head`, while the LLM decoder blocks and most CLIP/SAM backbone weights remain frozen.

**Stage 2 optimization.** Self-alignment initializes from the Stage-1 checkpoint and runs for 8 epochs with AdamW, weight decay 0.01, learning rate $5 \times 10^{-7}$, gradient accumulation 4, global gradient clipping 0.5, and a linear-warmup cosine-decay schedule with 30 warmup steps. Sampled captions use temperature 0.5 and nucleus sampling with top-$p = 0.9$, while the self-critical baseline caption is decoded greedily. Reward computation uses visibility threshold $\tau_{\text{vis}} = 0.3$, clipped-advantage threshold $\delta = 0.5$, and weights $(\beta, \gamma, \lambda_{\text{KL}}) = (0.15, 0.80, 0.20)$. Stage 2 updates only the language-side generator parameters (`model.layers`, `model.embed_tokens`, `model.norm`, and `lm_head`); the frozen reference model provides both reward-side `[DET]` states and the KL anchor, and all verifier modules remain frozen during reward evaluation.

# B Benchmark Protocols and Efficiency

**iGround.** We report the main iGround comparisons on the test set and mechanism ablations on the validation set, following GROVE. Both stages use videos from the iGround training split. Stage 1 additionally uses the annotated tagged captions and dense phrase tubes for supervised adaptation, whereas Stage 2 reuses the same training videos only: no caption, box, or tube annotations are consumed during self-alignment. Unless stated otherwise, validation decoding uses post-processing thresholds `conf=0.3` and `nms_iou=0.5`. iGround provides dense tube annotations, so it supports both trajectory-level Stage-1 matching and the full presence-plus-temporal Stage-2 reward. All quantitative evaluation is performed on the held-out iGround validation/test splits using their respective annotations.

**VidSTG.** VidSTG is evaluated under given-sentence grounding with ground-truth temporal localization. The query sentence is provided in teacher-forcing mode as in GROVE, so this benchmark evaluates transfer of the multi-instance grounding interface rather than caption refinement. The reported `Ours` row uses a VidSTG-finetuned checkpoint initialized from the HowToGround1M-pretrained GROVE weights rather than transferring the iGround checkpoint. Because the benchmark expects a single output tube, we collapse the $K$ slots after post-processing by retaining the highest-confidence slot; we never use oracle slot selection.

**ActivityNet-Entities.** ActivityNet-Entities is evaluated under sparse frame-level phrase grounding. Following GROVE, we use dataset-specific fine-tuning; since only sparse frame annotations are provided, temporal matching reduces to assignment on the annotated frame. For this benchmark we apply self-alignment with an objectness-derived sparse frame-level presence signal only, omitting the temporal-consistency reward. Thus the verifier acts as a support signal on annotated frames rather than as a dense temporal tube verifier, and full reward ablations are reported on iGround.

**Computational overhead.** The proposed multi-instance design remains efficient in practice. Under the same validation-time decoding and post-processing protocol, end-to-end inference latency remains

comparable to the GROVE baseline. On two NVIDIA A100 80GB GPUs, the multi-instance supervised adaptation completes in only 49 minutes, showing that the one-to-many upgrade is lightweight in wall-clock cost.

## C Additional Sensitivity Analyses

Table 5 clarifies the operating range of the design. Orthogonal initialization improves stability, but the more important pattern is the slot budget: increasing from a single slot to $K{=}3$ supplies the missing multi-instance capacity, whereas larger budgets bring little additional benefit and can dilute specialization. The KL sweep shows a similar balance on the language side. Too weak an anchor leaves the generator under-constrained, while too strong an anchor prevents useful visual feedback from changing the captioner; the default value gives the best tradeoff.

Table 5: iGround validation sensitivity to slot initialization, slot budget, and KL strength.

| | | Grounding | | Captioning | |
|---|---|---|---|---|---|
| Block | Variant | AP50 ↑ | Recall ↑ | METEOR ↑ | CIDEr ↑ |
| Multi-Instance Base sensitivity | random slot initialization | 51.4 | 27.8 | 19.4 | 92.9 |
| | reduced slot budget ($K = 2$) | 52.1 | 28.1 | 19.8 | 93.4 |
| | default slot budget ($K = 3$) | **52.7** | **28.4** | **19.9** | **93.8** |
| | larger slot budget ($K = 4$) | 52.4 | 28.3 | 19.8 | 93.5 |
| | larger slot budget ($K = 5$) | 52.6 | 28.2 | 19.8 | 93.3 |
| Visual Self-Alignment sensitivity | weak KL anchor ($\lambda_{\mathrm{KL}}{=}0.10$) | 53.4 | 29.1 | 20.2 | 96.4 |
| | default KL anchor ($\lambda_{\mathrm{KL}}{=}0.20$) | **53.9** | **29.7** | **20.5** | **97.1** |
| | strong KL anchor ($\lambda_{\mathrm{KL}}{=}0.50$) | 53.3 | 29.0 | 20.1 | 95.5 |
| | very strong KL anchor ($\lambda_{\mathrm{KL}}{=}1.00$) | 52.9 | 28.7 | 20.0 | 94.2 |

## D Limitations

Our method is designed as a lightweight upgrade to a strong pretrained grounded captioning stack, and therefore remains bounded by the visual and linguistic coverage of that stack. The multi-instance adapter exposes richer evidence than the original single-trajectory interface, but it is not intended to replace the underlying visual encoder, phrase-state extraction, or grounding decoder. Extremely subtle entities, severe occlusion, heavy motion blur, or scenes with more valid instances than the fixed slot budget $K$ may still require stronger backbones or adaptive slot allocation. These cases are natural extensions rather than contradictions of the proposed one-to-many interface.

The self-alignment stage deliberately uses a frozen verifier to keep the reward stable and annotation-free. This choice is useful for isolating visual feedback from additional supervision, but it also means that refinement is guided by the verifier's current visual evidence rather than by an independent oracle. KL regularization and advantage clipping help keep the generator close to the supervised reference, but future work could further improve robustness with uncertainty-aware rewards, adaptive verifier updates, or larger unlabeled video corpora.

Finally, our full closed-loop caption-refinement analysis is centered on iGround, the benchmark that provides the dense phrase grounding needed to evaluate both generation and grounding. VidSTG and ActivityNet-Entities provide complementary evidence under fixed-query and sparse-frame protocols, but they do not exercise every component of the full setting in the same way. Broader validation on longer videos, more open-vocabulary domains, and user-facing captioning scenarios remains an important next step.