CRTRACK-CAP: A NEW BENCHMARK FOR CROSS-VIEW DENSE VIDEO OBJECT CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense video object captioning (DVOC) localizes, tracks, and describes object trajectories within a video. Existing methods process each camera view independently, producing separate camera-local tracklets and captions for the same person across cameras. These outputs lack a unified description and leave complementary appearance evidence across views unused. We therefore formulate **cross-view DVOC (CVDVOC)**, which associates these tracklets into scene-level global trajectories and generates one caption for each trajectory from synchronized views. To support this task, we construct **CRTrack-Cap** from CRTrack by auditing text–identity links and converting usable referring expressions into canonical captions. It retains 425 global trajectories and 838 local tracklets across 13 scenes, with canonical captions for 335 identities. We adapt CHOTA to the cross-view setting as CV-CHOTA to jointly evaluate tracking and caption quality. Our reference framework combines **View-Exclusive Cross-Camera Association (VECA)** for grouping tracklets with **Conflict-Evidence Gating (CEG)** for integrating caption evidence using its reliability and disagreement across views. Against three single-view methods adapted with MvMHAT for cross-view association, our framework achieves CV-CHOTA scores of 83.43 and 33.48 on the scene-disjoint in-domain and cross-domain tests, respectively. Together, CRTrack-Cap and our framework provide a benchmark and a strong baseline for describing the same person consistently across camera views.

1 INTRODUCTION

Dense video object captioning (DVOC) combines object localization, tracking, and trajectory description within a video (Zhou et al., 2025; Fiastre et al., 2026). Existing methods process synchronized views independently, producing camera-local tracklets (per-camera trajectories) and separate captions for the same person. This leaves descriptions disconnected across cameras and complementary appearance evidence unused: *a bag hidden in one view may be visible in another.* Combining these views can therefore support a more complete and reliable global caption by incorporating appearance details that may be occluded in individual views. We formulate **cross-view DVOC (CVDVOC)**, which associates camera-local tracklets of the same person into a scene-level global trajectory and generates one caption from the combined visual evidence. Figure 1 illustrates this extension.

Evaluating this task requires synchronized videos, cross-camera identity annotations, and reference descriptions for global trajectories. BenSMOT supplies trajectory-level language for individual videos (Li et al., 2024), whereas DIVOTrack focuses on cross-camera correspondence (Hao et al., 2024). CRTrack provides synchronized videos, boxes, global identities, and referring expressions (Chen et al., 2025), making it a suitable starting point. However, its expressions specify tracking targets rather than provide reference captions for global trajectories. One expression may refer to several people, several expressions may refer to the same person, and some identities have no usable individual description. Direct reuse would therefore leave caption assignment to individual identities and eligibility for caption evaluation unclear.

We construct **CRTrack-Cap** by auditing text–identity links in both directions. For each identity, usable individual descriptions are normalized into a canonical appearance caption focused on stable, visible attributes, with compatible alternatives retained as references. Shared expressions retain their

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.

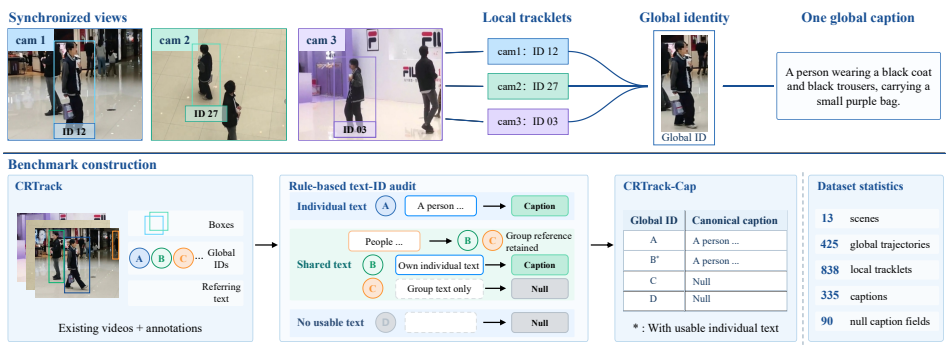


Figure 1: Cross-view DVOC and construction of CRTrack-Cap. Top: synchronized observations with different camera-local identifiers are linked into one global trajectory with one caption. Bottom: CRTrack videos, boxes, global identities, and referring expressions are converted into an auditable caption layer. Individual text supports a canonical caption for A and B; shared text retains its group scope and does not by itself supply a caption for C. D has no usable text. C and D retain their tracking annotations with null canonical-caption fields.

group scope rather than being automatically assigned as individual captions, and the source of each caption is recorded. Identities without usable individual text retain their tracking annotations but receive null captions. This preserves all identities for tracking evaluation while explicitly defining the subset eligible for caption evaluation. Across 13 scenes, CRTrack-Cap contains 425 global trajectories formed from 838 local tracklets, with canonical captions for 335 identities. We provide scene-disjoint in-domain and cross-domain tests and adapt CHOTA into CV-CHOTA for joint evaluation of tracking and caption quality at the global-trajectory level.

Cross-view DVOC requires deciding which observations belong to the same person and how they should contribute to one caption. A straightforward extension can associate camera-local tracklets and average their caption features, but appearance similarity alone does not enforce view exclusivity, and uniform averaging does not account for differences in evidence reliability. Our reference framework addresses these requirements with **View-Exclusive Cross-Camera Association (VECA)** and **Conflict-Evidence Gating (CEG)**. VECA merges cross-camera matches while allowing at most one tracklet from each camera in a group. Within each group, CEG weights semantic evidence and uses reliability and disagreement to regulate the caption-prompt update.

Our contributions are:

- We extend **single-view DVOC to cross-view DVOC**, generating one description for a global trajectory from observations across cameras. The task requires associating observations of the same person and combining their complementary appearance evidence.
- We construct **CRTrack-Cap** through an auditable mapping between referring expressions and global identities in 13 multi-camera scenes. It retains 425 identities for tracking and provides 335 canonical appearance captions, enabling joint evaluation on scene-disjoint in-domain and cross-domain tests.
- We provide a reference framework combining **VECA** for cross-view association and **CEG** for caption fusion. It achieves the highest CV-CHOTA among the compared methods on the in-domain and cross-domain tests, while component comparisons distinguish the contributions of association and caption generation.

2 Related Work

Dense Video Object Captioning. Dense video captioning localizes and describes events (Krishna et al., 2017; Wang et al., 2021; Yang et al., 2023), with recent extensions to streaming inference and event-duration imbalance (Zhou et al., 2024; Wu et al., 2025). Region-level approaches connect localization with language through dense-caption pretraining and region-to-text generation, as in CapDet and GRiT (Long et al., 2023; Wu et al., 2024); video localized narratives and grounded pretraining extend this connection over time (Voigtlaender et al., 2023; Kazakos et al., 2025). Complementary video-grounding methods, including STVGBert and TubeDETR (Su et al., 2021; Yang et al., 2022),

localize trajectories specified by input language. These tasks differ in whether language specifies or describes a target.

DVOC makes the object trajectory the description unit. DenseVOC jointly detects, tracks, and captions objects from disjoint supervision (Zhou et al., 2025); BenSMOT provides semantic tracking annotations, and SMOTer predicts trajectories and their descriptions (Li et al., 2024). OW-VISCapTor extends instance segmentation with open-world captioning (Choudhuri et al., 2024), and CaptionFormer unifies segmentation, tracking, and trajectory captioning (Fiastre et al., 2026). CRTrack-Cap extends this camera-local unit to global trajectories, combining observations of one identity across cameras.

Cross-View Multi-Object Tracking. Single-camera MOT supplies local trajectories and appearance features for cross-view processing. Detection-based methods include ByteTrack and UCMCTrack (Zhang et al., 2022b; Yi et al., 2024); FairMOT and RelationTrack jointly learn detection and appearance (Zhang et al., 2021; Yu et al., 2023). Transformer-based methods model temporal relations or propagate track queries, as in Global Tracking Transformers, MOTR, and MOTRv2 (Zhou et al., 2022; Zeng et al., 2022; Zhang et al., 2023). Recent work explores identity prediction, learnable association, and discriminative embedding refinement (Gao et al., 2025; Wang et al., 2025b; Shao et al., 2026). These methods provide the camera-local inputs that must be associated before caption evidence can be combined across views.

Across cameras, temporal continuity alone no longer resolves identity. Deep assignment networks model relations between observations (Han et al., 2022). MvMHAT uses self-supervised multi-view association (Gan et al., 2021), and MvMHAT++ further studies spatial–temporal consistency and cross-view evaluation (Feng et al., 2025). DIVOTrack, All-Day MCMT, and GMT develop cross-view benchmarks or association systems under diverse conditions (Hao et al., 2024; Fan et al., 2025; Zhen et al., 2026); MITracker studies multi-view integration in the related visual object tracking setting (Xu et al., 2025). Appearance representations face additional variation from clothing changes and synthetic-to-real transfer, as studied in video person Re-ID (Wang et al., 2025a; Zhang et al., 2025).

Language introduces a further distinction. Open-vocabulary tracking broadens the tracked category set, with COVTrack integrating semantic, motion, and appearance cues (Qian et al., 2025). Referring tracking instead uses language to select targets: iKUN connects text to existing trackers (Du et al., 2024), CRMOT extends this problem across cameras (Chen et al., 2025), and CD-RMOT-Bench studies cross-domain transfer (Zhang et al., 2026). Cross-view DVOC instead extends caption generation across cameras, using association to collect observations of the same target.

Evaluation of Tracking and Captioning. Tracking benchmarks assess localization and identity continuity (Dendorfer et al., 2021; Ristani et al., 2016); HOTA separates detection and association accuracy (Luiten et al., 2021), and multi-view evaluation additionally measures cross-camera identity consistency (Feng et al., 2025; Chen et al., 2025). Caption metrics assess different aspects of language agreement: CIDEr measures consensus, METEOR lexical matching, and SPICE semantic structure (Vedantam et al., 2015; Banerjee & Lavie, 2005; Anderson et al., 2016). DenseVOC introduced CHOTA to couple detection, association, and caption accuracy, and CaptionFormer adopted it for DVOC (Zhou et al., 2025; Fiastre et al., 2026). We evaluate global trajectories with CV-CHOTA and its tracking and language components.

3 CRTrack-Cap Benchmark

3.1 Task and Prediction Unit

Given synchronized videos $\mathcal{V} = \{V^k\}_{k=1}^{K}$, cross-view DVOC predicts global trajectories and one caption per trajectory. Our framework first extracts camera-local tracklets with identifiers, frame indices, and boxes. Each tracklet is augmented with two features that serve different downstream roles: a CaptionFormer trajectory query for language generation and an appearance descriptor for association. The resulting common record is

$$\mathcal{T} = \{T_i^k\}, \qquad T_i^k = (k, \ell_i^k, \mathcal{F}_i^k, \mathcal{B}_i^k, \mathbf{q}_i^k, \mathbf{e}_i^k). \tag{1}$$

Here i indexes tracklets within camera k; ℓ_i^k is the local identifier, $\mathcal{F}_i^k$ is the set of observed frames, and $\mathcal{B}_i^k$ contains boxes in $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ format. A local tracklet is a camera-specific

temporal unit before cross-view association. The query $\mathbf{q}_i^k \in \mathbb{R}^{d_q}$ carries the visual representation consumed by the caption pathway, whereas $\mathbf{e}_i^k \in \mathbb{R}^{d_e}$ represents appearance for cross-camera matching.

A cross-view system partitions the local tracklets into global identity groups $\widehat{\mathcal{G}} = \{\widehat{G}_n\}$. A scene-qualified identity and its time-stamped member tracklets constitute a global trajectory; the members retain their native camera coordinates. The system predicts one global caption $\widehat{C}_n$ per group:

$$\widehat{G}_n \subseteq \mathcal{T}, \qquad \widehat{Y}_n = (\widehat{G}_n, \widehat{C}_n). \tag{2}$$

Here n indexes predicted groups, which are disjoint and cover $\mathcal{T}$; $\widehat{Y}_n$ is the global-trajectory-and-caption prediction. A ground-truth key $g = (\text{scene}, \text{id})$ distinguishes identities across scenes. A missing reference caption is denoted by $c_g = \varnothing$; the identity remains in tracking evaluation but is excluded from caption supervision.

3.2 An Auditable Caption Layer

CRTrack supplies referring expressions rather than a one-caption-per-global-trajectory annotation. Let $\mathcal{E}$ denote its expressions and $\mathcal{I}$ the scene-qualified ground-truth global identities. We represent its supervision as a relation $R \subseteq \mathcal{E} \times \mathcal{I}$ and retain both directions:

$$\Phi_E(e) = \{g : (e, g) \in R\}, \qquad \Phi_G(g) = \{e : (e, g) \in R\}. \tag{3}$$

The two relation maps expose four cases: unique mappings, multiple expressions for one identity, expressions shared by several identities, and identities without text. Figure 1 illustrates their treatment. Unique candidates undergo light normalization that preserves visible attributes. Compatible alternatives are retained as references, with one selected as the canonical caption. Conflicting descriptions remain traceable through the review record. Shared expressions retain their group scope: B's additional individual text supports its canonical caption, whereas C's shared reference alone does not. Identities without usable individual language receive a null canonical field and retain their local tracklets and global identity.

Converted samples retain source expressions, relation maps, coordinates, and frame paths. CRTrack-Cap uses these records to evaluate cross-view identity association and captions of stable, visible appearance, while reusing the released boxes and cross-camera identities.

3.3 Dataset Composition and Splits

We use scene-disjoint training, validation, in-domain test, and cross-domain test splits based on CRTrack (Chen et al., 2025) (Appendix A.2). Training and in-domain test scenes come from DIVOTrack's moving-camera videos; cross-domain test scenes come from CAMPUS's static-camera videos. The latter therefore evaluate transfer across source domains and capture conditions. Tracking evaluation retains all identities, while caption supervision and scoring use those with canonical descriptions. Table 1 reports both populations.

Table 1: CRTrack-Cap statistics. Coverage is the percentage of global trajectories with a canonical caption.

Split	Scenes	Identities	Captioned	Missing	Coverage
Train	6	256	196	60	76.56
Validation	1	33	14	19	42.42
In-domain test	3	87	76	11	87.36
Cross-domain test	3	49	49	0	100.00
Total	13	425	335	90	78.82

The text–identity audit identifies 341 one-text/one-identity candidates, three multiple-text/one-identity cases, and 81 identities without a source expression. Nine additional candidates lack usable canonical text. These counts preserve every global identity while defining the 335-identity language subset. Serialization, coordinate, and local-to-global checks are summarized in Appendix A.

3.4 Scene-level Tracking and Caption Evaluation

Synchronized camera sequences form a scene-level pseudo-sequence with global identity labels. TrackEval computes CVIDF1 (global identity F1), CV-AssA (association accuracy), and CV-HOTA (joint detection and association accuracy) on this representation (Feng et al., 2025). CV-HOTA and CV-AssA average over the standard intersection-over-union threshold grid; the caption-aware joint metric uses detection and association accuracy at threshold 0.50. Appendix A.6 defines the tracking metrics.

For language evaluation, Hungarian matching on HOTA's identity-alignment matrix pairs predicted groups with ground-truth identities. For identity g, let $m(g)$ denote its matched group, if any, and define $\hat{c}_g = \widehat{C}_{m(g)}$ when matched and an empty string otherwise. Each caption-eligible identity contributes once; unmatched predicted groups affect only tracking terms. Following CaptionFormer's scorer combination, we define

$$\text{Global-CapA} = \frac{1}{3|\mathcal{G}_{\text{cap}}|} \sum_{g \in \mathcal{G}_{\text{cap}}} \big[\text{CIDEr}_{\mathcal{D}}(c_g, \hat{c}_g) + \text{METEOR}_{\mathcal{D}}(c_g, \hat{c}_g) + \text{SPICE}_{\mathcal{D}}(c_g, \hat{c}_g)\big]. \tag{4}$$

Here $\mathcal{G}_{\text{cap}}$ contains all scene-qualified ground-truth identities with a canonical caption c_g, and $\hat{c}_g$ is the aligned prediction (empty if unaligned). Each scorer processes the full indexed pair list $\mathcal{D} = ((c_g, \hat{c}_g))_{g \in \mathcal{G}_{\text{cap}}}$ once; the subscript denotes its returned per-identity scores, including CIDEr's corpus-dependent statistics (Vedantam et al., 2015).

The joint score combines detection, global association, and caption accuracy:

$$\text{CV-CHOTA}_{0.50} = \big(\text{DetA}_{0.50} \cdot \text{CV-AssA}_{0.50} \cdot \text{Global-CapA}\big)^{1/3}. \tag{5}$$

All caption components retain their native pycocoevalcap scales. CIDEr can exceed one; consequently, Global-CapA and CV-CHOTA need not lie in $[0, 1]$. In tables, CV-CHOTA uses the 0.50 threshold and all metrics are displayed on the $\times 100$ scale. Reporting the components separates language from association changes.

4 Reference Framework

Our framework follows the task decomposition in Figure 2. A local front end exports tracklet geometry, an appearance descriptor for cross-camera association, and a CaptionFormer query for language generation. VECA forms view-exclusive global identity groups, and CEG uses query-derived support and disagreement to adapt their caption prompts.

4.1 View-Exclusive Cross-Camera Association

VECA addresses identity fragmentation over the fixed local tracklets. For tracklets in different cameras, it normalizes appearance descriptors and computes their similarity:

$$\bar{\mathbf{e}}_i^k = \frac{\mathbf{e}_i^k}{\|\mathbf{e}_i^k\|_2}, \qquad S_{ij}^{ku} = \langle \bar{\mathbf{e}}_i^k, \bar{\mathbf{e}}_j^u \rangle, \qquad k \neq u. \tag{6}$$

For each unordered camera pair, Hungarian assignment maximizes total similarity. Assigned edges are retained when $S_{ij}^{ku} \geq \tau$, with $\tau = 0.54$ calibrated by pairwise association F1 on training annotations. Pairwise optimality is insufficient for a valid global identity group because transitive links can place two tracklets from one camera in the same component. VECA therefore processes retained edges in descending score order and merges components A and B only if

$$\mathcal{C}(A) \cap \mathcal{C}(B) = \varnothing. \tag{7}$$

Here $\mathcal{C}(A)$ is the set of cameras represented in A. Tracklets without accepted links remain singleton global identity groups. This merge enforces VECA's one-member-per-camera assumption, $|\widehat{G}_n \cap \mathcal{T}^k| \leq 1$, where $\mathcal{T}^k$ contains the tracklets from camera k. VECA returns both group membership and the original member tracklets, allowing the caption pathway to inspect every contributing view.

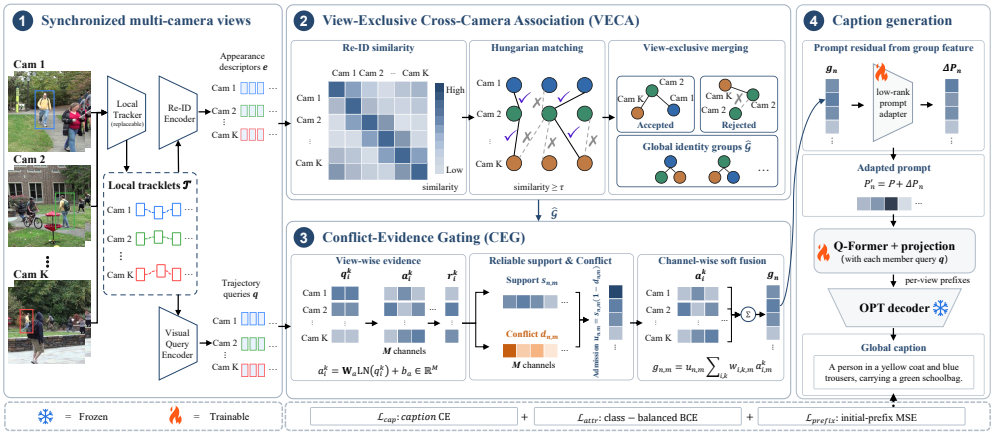


Figure 2: Reference framework for cross-view DVOC. The local tracker exports tracklets, Re-ID descriptors $\mathbf{e}$, and trajectory queries $\mathbf{q}$. VECA forms global identity groups; CEG combines member evidence with reliability and conflict estimates to produce a prompt residual ΔP_n. The Q-Former processes each member query with this adapted prompt, and projected prefixes are averaged before frozen OPT generates one caption. We use CRTracker as the local tracker. The bottom row shows the training losses in Equation 14.

4.2 Conflict-Evidence Gating

For each member query in a VECA group, CEG uses a scalar reliability from alignment and tracklet geometry:

$$r_i^k = \max(J_i^k, 0.05)\sqrt{A_i^k}\,\min\left(1, \frac{\log(1+L_i^k)}{\log 21}\right). \tag{8}$$

Here $J_i^k \in [0,1]$ is the cached mask–tracklet-box IoU, $A_i^k \in [10^{-6}, 1]$ is representative box area normalized by the padded, resized image area and clipped to that interval, and L_i^k is tracklet length. Larger aligned observations receive more weight, while duration support saturates at 20 frames. This geometric visibility proxy is computed independently of the semantic logits; the same scalar accompanies all evidence channels of that tracklet.

MVMean averages member language-prefix matrices after the query adapter, Q-Former, and projection. We denote this complete member pathway by $\mathrm{CFPrefix}(\mathbf{q}; P)$, with learned prompt P (Appendix A.4).

CEG weights members separately for each evidence channel. A channel head maps each normalized member query into logits:

$$\mathbf{a}_i^k = \mathbf{W}_a\,\mathrm{LN}(\mathbf{q}_i^k) + \mathbf{b}_a, \qquad \mathbf{a}_i^k \in \mathbb{R}^M, \quad \mathbf{W}_a \in \mathbb{R}^{M\times d_q}. \tag{9}$$

The $M = 88$ channels provide structured weak supervision over stable appearance attributes. Each channel predicts a weak label obtained by deterministic parsing of the canonical training captions. An unmentioned attribute is coded as zero for this supervision, although silence in a caption does not establish visual absence. A class-balanced binary cross-entropy term with weight 0.05 supervises the fused group logits before the admission gate (Appendix A.4). For channel m, let $p_{i,m}^k = \mathrm{sigmoid}(a_{i,m}^k)$ and $\chi_{i,m}^k = |2p_{i,m}^k - 1|$. The first quantity predicts the caption-derived label, while the second measures confidence away from the ambiguous midpoint.

Let $\mathcal{I}_n = \{(i,k) : T_i^k \in \widehat{G}_n\}$ index group members. All sums, maxima, and softmax operations below run over $\mathcal{I}_n$, separately for each channel m. The normalized reliabilities are

$$\pi_i^k = \frac{r_i^k}{\sum_{(j,v)\in\mathcal{I}_n} r_j^v}, \qquad \hat{r}_i^k = \frac{r_i^k}{\max_{(j,v)\in\mathcal{I}_n} r_j^v}. \tag{10}$$

With the training-set reliability percentile $r_{0.9}$ as an absolute-quality reference, support and conflict are

$$
\begin{aligned}
s_{n,m} &= \mathrm{clip}\left(\frac{\max_{(i,k)\in\mathcal{I}_n} r_i^k}{r_{0.9}}, 0, 1\right) \max_{(i,k)\in\mathcal{I}_n} \left(\hat{r}_i^k \chi_{i,m}^k\right), \\
d_{n,m} &= \mathrm{clip}\left(4 \sum_{(i,k)\in\mathcal{I}_n} \pi_i^k (p_{i,m}^k - \bar{p}_{n,m})^2, 0, 1\right), \\
\bar{p}_{n,m} &= \sum_{(i,k)\in\mathcal{I}_n} \pi_i^k p_{i,m}^k.
\end{aligned}
\tag{11}
$$

Support measures reliable certainty; conflict measures probability disagreement. Their product $u_{n,m} = s_{n,m}(1 - d_{n,m})$ admits evidence that is both supported and cross-view coherent. CEG then performs channel-specific soft fusion:

$$
\begin{aligned}
w_{i,k,m} &= \mathrm{softmax}_{(i,k)\in\mathcal{I}_n} \left(\log \max(r_i^k, \epsilon) + a_{i,m}^k\right), \\
z_{n,m} &= \sum_{(i,k)\in\mathcal{I}_n} w_{i,k,m} a_{i,m}^k, \qquad g_{n,m} = u_{n,m} z_{n,m}.
\end{aligned}
\tag{12}
$$

Here ϵ is the floating-point machine epsilon used to clamp reliability before taking the logarithm. The softmax weights select the members that contribute to channel m, while $u_{n,m}$ determines how strongly their fused evidence can modify the caption prompt. This separates evidence selection from conflict suppression.

A rank-8 adapter maps $\mathbf{g}_n \in \mathbb{R}^M$ to a residual $\Delta P_n \in \mathbb{R}^{N_q \times d_p}$ for the N_q learned Q-Former tokens of width d_p:

$$
\begin{aligned}
P_n' &= P + \Delta P_n, \\
\mathbf{p}_n &= \frac{1}{|\mathcal{I}_n|} \sum_{(i,k)\in\mathcal{I}_n} \mathrm{CFPrefix}(\mathbf{q}_i^k; P_n').
\end{aligned}
\tag{13}
$$

Here ρ_n is the mean admission across channels, and $0.1\rho_n$ scales the residual (Appendix A.4). Members share the adapted prompt; their averaged prefix conditions the frozen OPT decoder (Zhang et al., 2022a). The query adapter, Q-Former (including P), and projections form the 107.36M trainable caption pathway; CEG adds 245,200 parameters. The visual encoder, tracker, Re-ID encoder, and OPT remain frozen, and VECA is fixed.

4.3 Training Objective

The caption pathway learns to generate the canonical description while retaining the pretrained prefix representation:

$$
\mathcal{L} = \mathcal{L}_{\mathrm{cap}} + 0.05\, \mathcal{L}_{\mathrm{attr}} + \mathcal{L}_{\mathrm{prefix}}. \tag{14}
$$

Here $\mathcal{L}_{\mathrm{cap}}$ is teacher-forced caption cross-entropy, $\mathcal{L}_{\mathrm{attr}}$ is class-balanced binary cross-entropy on the pre-admission logits $z_{n,m}$ with caption-derived weak labels, and $\mathcal{L}_{\mathrm{prefix}}$ is mean-squared error between the group prefix $\mathbf{p}_n$ and its initial CaptionFormer counterpart. The weak-label term grounds evidence channels in the training captions; the prefix term anchors adaptation to the pretrained language interface. Appendix A.4 specifies optimization details.

5 Experiments

5.1 Experimental Setup

Comparison protocol. To our knowledge, no existing method directly addresses cross-view dense video object captioning (CVDVOC). We therefore adapt three single-view comparison methods capable of trajectory-level description—DenseVOC (Zhou et al., 2025), SMOTer (Li et al., 2024), and CaptionFormer (Fiastre et al., 2026). After fine-tuning their native tracking components, we apply the same MvMHAT (Gan et al., 2021) algorithm to associate their local trajectories across synchronized views. We evaluate the resulting global trajectories and captions on the same CRTrack-Cap splits with a common evaluator. Our framework uses VECA for cross-view association and CEG for caption fusion; Table 3 compares the component configurations on fixed local predictions.

Training and evaluation. For caption training, local tracklets and their assigned global identity groups are fixed. We optimize the 107.36M-parameter caption pathway and the 245K-parameter CEG adapter, using one caption-eligible global identity per step. Training runs for 70 epochs with AdamW, learning rates of 10^{-6} and 10^{-4} for the caption pathway and CEG, respectively, weight decay 10^{-2}, and gradient clipping at one. We report full-split results on the in-domain and cross-domain tests.

5.2 Scene-level Task Performance

Table 2: Main comparison on CRTrack-Cap with predicted local tracklets. DenseVOC, SMOTer, and CaptionFormer use MvMHAT for cross-view association; ours uses VECA. Metric values are shown on the $\times 100$ scale; higher is better.

Method	CVIDF1	CV-HOTA	CV-AssA	CV-CHOTA	Global-CapA	CIDEr
In-domain test						
DenseVOC	66.28	55.79	51.64	48.17	24.63	26.70
SMOTer	59.83	51.80	44.82	57.43	48.18	66.39
CaptionFormer	77.35	64.28	58.44	76.34	75.54	133.53
Ours	**80.65**	**68.57**	**62.84**	**83.43**	**88.19**	**157.04**
Cross-domain test						
DenseVOC	15.69	16.15	5.98	19.58	19.11	17.14
SMOTer	**18.87**	**20.88**	**8.22**	30.82	41.45	63.36
CaptionFormer	16.40	16.87	6.38	29.94	60.63	97.70
Ours	17.00	18.71	7.27	**33.48**	**68.34**	**112.62**

Table 2 compares complete pipelines with different local trackers, caption pathways, and association rules. Our framework leads all in-domain metrics. Cross-domain tracking is strongest for SMOTer, but our framework leads CV-CHOTA, Global-CapA, and CIDEr, showing that tracking strength alone does not determine global-caption quality. The component study separates association from caption fusion.

5.3 Separating Association from Caption Fusion

Table 3 compares a 2×2 grid using the same predicted local tracklets and evaluator. The association configurations are singleton groups and VECA groups; the caption configurations use separately trained MVMean and CEG pathways. The VECA+CEG rows report the main-table results.

Table 3: Association and caption-fusion configurations; metric values are shown on the $\times 100$ scale. The VECA+CEG rows reproduce Table 2; the other rows are the recorded component controls.

Setting	VECA	CEG	CVIDF1	CV-HOTA	CV-AssA	CV-CHOTA	Global-CapA	CIDEr
In-domain test								
Singleton + MVMean			50.65	51.86	35.96	67.04	80.19	138.03
Singleton + CEG		✓	50.65	51.86	35.96	68.56	85.77	149.01
VECA + MVMean	✓		**80.65**	**68.57**	**62.84**	79.92	77.52	130.20
VECA + CEG	✓	✓	**80.65**	**68.57**	**62.84**	**83.43**	**88.19**	**157.04**
Cross-domain test								
Singleton + MVMean			14.87	17.34	6.25	29.83	56.37	87.86
Singleton + CEG		✓	14.87	17.34	6.25	29.54	54.73	82.67
VECA + MVMean	✓		**17.00**	**18.71**	**7.27**	31.45	56.64	82.46
VECA + CEG	✓	✓	**17.00**	**18.71**	**7.27**	**33.48**	**68.34**	**112.62**

VECA raises in-domain CVIDF1 by 30.00 points and CV-AssA by 26.88 over singleton groups, but MVMean lowers Global-CapA after this grouping change. With VECA fixed, CEG raises Global-CapA from 77.52 to 88.19 in-domain and from 56.64 to 68.34 cross-domain, while tracking metrics remain unchanged. Thus, association supplies the global groups, whereas evidence-aware fusion determines how those groups improve captions. CEG lowers caption scores for singleton groups but improves them with VECA groups fixed, and VECA+CEG attains the grid's highest joint and caption scores on both splits.

The recorded association-rule comparison gives 80.65 in-domain CVIDF1 for full VECA versus 78.22 for greedy view-exclusive matching. Unconstrained graph closure creates 5 in-domain and 46 cross-domain same-camera collisions; view-exclusive groups avoid these invalid links (Appendix A.3).

View-count analysis. Table 4 stratifies the VECA+CEG and VECA+MVMean configurations from Table 3 by the number of cameras observing each caption-eligible reference identity.

Table 4: View-count comparison with VECA groups fixed. Metric deltas are CEG minus MVMean on the $\times 100$ scale. Sample counts and additional metrics appear in Appendix A.5.

Split	Views	ΔCV-CHOTA	ΔCIDEr
In-domain	1	+0.11	+0.89
In-domain	2	+6.55	+55.03
In-domain	3	+4.11	+38.48
Cross-domain	3	+2.22	+35.21
Cross-domain	4	+1.88	+29.13

The single-view in-domain group changes little, whereas both multi-view groups improve in CIDEr and CV-CHOTA; the two cross-domain groups also gain. This pattern supports the role of CEG in combining evidence across views, while the non-monotonic gains do not imply that adding cameras always yields larger improvements. Together, these results indicate that cross-view caption quality depends on both correct grouping and selective use of complementary evidence.

5.4 Qualitative Comparison

Figure 3 compares in-domain and cross-domain targets. Our framework and CaptionFormer link all shown views, but our captions better match the reference attributes: white shoes in the former and a yellow coat and schoolbag in the latter.

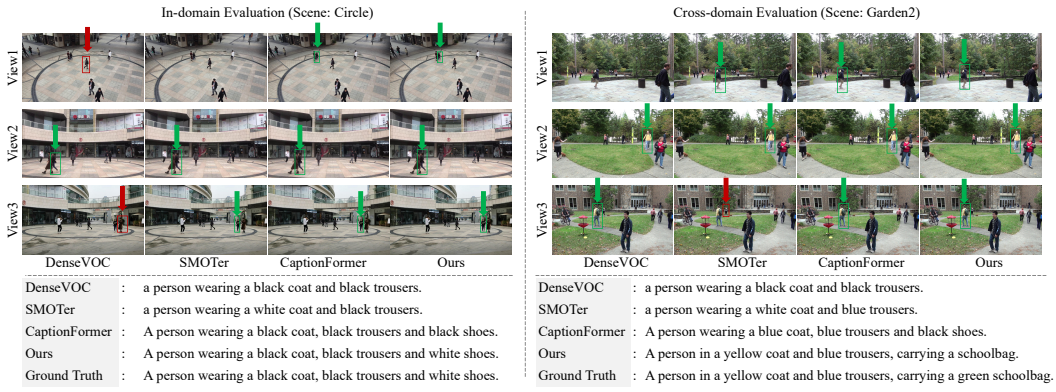


Figure 3: In-domain (Circle) and cross-domain (Garden2) examples. Rows are synchronized views; columns are methods. Green/red arrows mark correct/incorrect links; no arrow means unmatched. Text below compares generated captions with ground truth.

6 Conclusion

We extended single-view DVOC to captioning global trajectories across cameras. CRTrack-Cap supports this task with 425 tracking identities and 335 canonical captions, audited from referring expressions and evaluated on scene-disjoint in-domain and cross-domain tests. Our reference framework combines cross-view association with caption fusion and achieves the highest CV-CHOTA among the compared methods on both tests. Component comparisons distinguish association and caption gains under fixed local predictions. These results support evaluating association and caption quality jointly. They also show that the caption gains depend on how the associated evidence is fused, rather than on grouping alone. Cross-domain performance remains lower, motivating further study of captioning across capture conditions.

AI USE STATEMENT

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

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, 2016.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005.

Sijia Chen, En Yu, and Wenbing Tao. Cross-view referring multi-object tracking. In *AAAI*, 2025.

Anwesa Choudhuri, Girish Chowdhary, and Alexander Schwing. OW-VISCapTor: Abstractors for open-world video instance segmentation and captioning. In *NeurIPS*, 2024.

Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. MOTChallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 2021.

Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. IKUN: Speak to trackers without retraining. In *CVPR*, 2024.

Huijie Fan, Yu Qiao, Yihao Zhen, Tinghui Zhao, Baojie Fan, and Qiang Wang. All-day multi-camera multi-target tracking. In *CVPR*, 2025.

Wei Feng, Feifan Wang, Ruize Han, Yiyang Gan, Zekun Qian, Junhui Hou, and Song Wang. Unveiling the power of self-supervision for multi-view multi-human association and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):351–368, 2025.

Gabriel Fiastre, Antoine Yang, and Cordelia Schmid. CaptionFormer: Unified segmentation, tracking, and captioning for spatio-temporal objects. In *CVPR*, 2026.

Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *ACM MM*, 2021.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as ID prediction. In *CVPR*, 2025.

Ruize Han, Yun Wang, Haomin Yan, Wei Feng, and Song Wang. Multi-view multi-human association with deep assignment network. *IEEE Transactions on Image Processing*, 31:1830–1840, 2022.

Shengyu Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. DIVOTrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *International Journal of Computer Vision*, 2024.

Evangelos Kazakos, Cordelia Schmid, and Josef Sivic. Large-scale pre-training for grounded video caption generation. In *ICCV*, 2025.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

Yunhao Li, Qin Li, Hao Wang, Xue Ma, Jiali Yao, Shaohua Dong, Heng Fan, and Libo Zhang. Beyond MOT: Semantic multi-object tracking. In *ECCV*, 2024.

Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. CapDet: Unifying dense captioning and open-world detection pretraining. In *CVPR*, 2023.

Jonathon Luiten, Aljoša Ošep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 2021.

Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. COVTrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In *ICCV*, 2025.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016.

Yuqing Shao, Yuchen Yang, Rui Yu, Weilong Li, Xu Guo, Huaicheng Yan, Wei Wang, and Xiao Sun. From detection to association: Learning discriminative object embeddings for multi-object tracking. In *CVPR*, 2026.

Rui Su, Qian Yu, and Dong Xu. STVGBert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, 2021.

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

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives. In *CVPR*, 2023.

Likai Wang, Xiangqun Zhang, Ruize Han, Yanjie Wei, Song Wang, and Wei Feng. A new benchmark and algorithm for clothes-changing video person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:1993–2005, 2025a.

Peng Wang, Yongcai Wang, Hualong Cao, Wang Chen, and Deying Li. LA-MOTR: End-to-end multi-object tracking by learnable association. In *ICCV*, 2025b.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, 2021.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. GRiT: A generative region-to-text transformer for object understanding. In *ECCV*, 2024.

Kangyi Wu, Pengna Li, Jingwen Fu, Yizhe Li, Yang Wu, Yuhan Liu, Jinjun Wang, and Sanping Zhou. Event-equalized dense video captioning. In *CVPR*, 2025.

Mengjie Xu, Yitao Zhu, Haotian Jiang, Jiaming Li, Zhenrong Shen, Sheng Wang, Haolin Huang, Xinyu Wang, Han Zhang, Qing Yang, and Qian Wang. MITracker: Multi-view integration for visual object tracking. In *CVPR*, 2025.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-temporal video grounding with transformers. In *CVPR*, 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 *CVPR*, 2023.

Kefu Yi, Kai Luo, Xiaolei Luo, Jiangui Huang, Hao Wu, Rongdong Hu, and Wei Hao. UCMCTrack: Multi-object tracking with uniform camera motion compensation. In *AAAI*, 2024.

En Yu, Zhuoling Li, Shoudong Han, and Hongwei Wang. RelationTrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 2023.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

Xiangqun Zhang, Ruize Han, Likai Wang, Linqi Song, Junhui Hou, and Wei Feng. Synthetic-to-real video person Re-ID. *IEEE Transactions on Information Forensics and Security*, 20:5438–5450, 2025.

Xiangqun Zhang, Likai Wang, Zekun Qian, Ruize Han, and Wei Feng. CD-RMOT-Bench: Benchmarking the cross-domain referring multi-object tracking. *arXiv preprint arXiv:2607.25239*, 2026.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 2021.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022b.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *CVPR*, 2023.

Yihao Zhen, Mingyue Xu, Qiang Wang, Baojie Fan, Jiahua Dong, Tinghui Zhao, and Huijie Fan. GMT: Effective global framework for multi-camera multi-target tracking. In *CVPR*, 2026.

Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krahenbuhl. Global tracking transformers. In *CVPR*, 2022.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *CVPR*, 2024.

Xingyi Zhou, Anurag Arnab, Chen Sun, and Cordelia Schmid. Dense video object captioning from disjoint supervision. In *ICLR*, 2025.

A Additional Analyses and Reproducibility Details

This appendix follows the evaluated pipeline: caption-layer construction, cross-view association, CEG implementation, caption-evidence diagnostics, and metric computation.

A.1 Benchmark Construction and Source Audit

The index retains all 425 scene-qualified global trajectories and 838 local tracklets, including the 90 identities with null captions. Of the 425 identity records, 341 are one-text/one-identity candidates, three contain multiple expressions for one identity, and 81 have no source expression. Nine additional candidates lack usable canonical text. The release therefore contains a 425-identity tracking index and a 335-identity language-eligible subset.

Source validation records 4,936 border-clipped boxes, 11 sequence-dimension mismatches, and 138 annotation rows with unavailable frame images. These diagnostics are retained alongside source coordinates and image paths. Converted samples serialize without errors, and the local-to-global mapping check reports no conflicts. Conversion validity and source-frame availability are recorded as separate fields, preserving all missing-frame outcomes for audit.

A.2 Scene Splits and Domain Shift

The split manifest uses Floor, Gate1, Ground, Moving, Park, and Shop for training; Square for validation; Circle, Gate2, and Side for in-domain testing; and Garden1, Garden2, and ParkingLot for cross-domain testing. We retain CRMOT's in-domain and cross-domain test scenes and hold Square out from its original training scenes for development.

The domain distinction follows the source collections in CRTrack (Chen et al., 2025). Training and in-domain scenes originate from DIVOTrack, whereas the cross-domain scenes originate from CAMPUS. DIVOTrack uses three moving cameras; CAMPUS uses three or four static cameras. The change in capture setup, scene appearance, viewpoint, and trajectory duration makes cross-domain evaluation a transfer across source domains rather than only a new scene from the training collection.

A.3 Association-Rule Analysis

Table 5 compares the recorded association-rule settings on the in-domain test. Full VECA scores 80.65/68.57/62.84 CVIDF1/CV-HOTA/CV-AssA, compared with 78.22/67.75/61.18 for greedy view-exclusive matching. Their CV-DetA scores are close (74.99 versus 75.02). Ordinary graph closure produces 5 in-domain and 46 cross-domain same-camera collisions; view-exclusive merging avoids these invalid groups.

Table 5: Recorded association-rule comparison on the in-domain test. TrackEval metric values are shown on the $\times 100$ scale; best values are bold.

Rule	CVIDF1	CV-HOTA	CV-DetA	CV-AssA	CV-AssRe
Singleton	50.65	51.86	74.99	35.96	37.09
Greedy view-exclusive	78.22	67.75	**75.02**	61.18	63.79
Hungarian star	77.03	66.65	74.99	59.38	61.74
Full VECA	**80.65**	**68.57**	74.99	**62.84**	**65.39**

For the same in-domain greedy view-exclusive run, CV-CHOTA@0.50, Global-CapA, and CIDEr are 80.00, 82.74, and 142.62, respectively.

A.4 CEG Supervision and Optimization

The implementation uses $M = 88$ structured weak-label channels: 22 item-type atoms and 66 slot–color atoms. The latter factor six semantic slots with eleven colors. Canonical training captions are parsed deterministically into multi-label targets. The channel logits are explicitly supervised; attribute-level caption F1 measures attributes extracted from the final decoded captions.

The geometric reliability is defined in Equation 8. Its reference percentile pools member reliabilities from caption-eligible training identities, with $r_{0.9} = 0.08719$. Reliability supplies one scalar per local tracklet and is broadcast across the evidence channels. The query cache selects the predicted mask with maximum IoU against the supplied representative tracklet box rasterized at mask resolution; mask probabilities are thresholded at 0.5. The box is scaled and normalized by the padded image dimensions before computing its area fraction. In the predicted-tracklet pipeline, the supplied box comes from the local tracker.

Optimization combines teacher-forced caption cross-entropy, class-balanced weak-label binary cross-entropy with weight 0.05, and mean-squared error to the initial CaptionFormer prefix with weight 1.0. The residual bottleneck has rank eight, its scale is 0.1, and the up-projection is zero-initialized. The trainable caption path comprises a 33,600-parameter query adapter, approximately 105.16M Q-Former/query-token parameters, and 2,166,016 projection parameters.

Each optimization step uses one identity. Training uses no gradient accumulation, learning-rate scheduler, or warmup. The frozen OPT decoder uses half precision, while the optimizer runs without an automatic mixed-precision gradient scaler. Training uses one RTX 3090.

MVMean averages aligned language-prefix matrices from the shared query adapter, Q-Former, and projection, denoted $\mathrm{CFPrefix}(\mathbf{q}; P)$, where P is the learned Q-Former prompt:

$$\mathbf{p}_n^{\mathrm{mv}} = \frac{1}{|\widehat{G}_n|} \sum_{T_i^k \in \widehat{G}_n} \mathrm{CFPrefix}(\mathbf{q}_i^k; P). \tag{15}$$

Member indices and prompt mapping. The pair $(i, k) \in \mathcal{I}_n$ identifies a tracklet, not an independent camera–tracklet Cartesian product. Several tracklets may exist in a camera before association; VECA permits at most one in each predicted group. The head uses $W_a \in \mathbb{R}^{M \times d_q}$ and $\mathbf{b}_a \in \mathbb{R}^M$. The query $\mathbf{q}_i^k$ enters the evidence head before the caption query adapter. Appearance descriptors $\mathbf{e}_i^k$ enter VECA only. Layer normalization acts on the feature dimension; the fusion softmax normalizes across members for each fixed channel m, so $\sum_{(i,k)\in\mathcal{I}_n} w_{i,k,m} = 1$.

For $\mathbf{g}_n = (g_{n,1}, \ldots, g_{n,M})$, the prompt adapter is

$$\begin{aligned} \rho_n &= \frac{1}{M} \sum_{m=1}^{M} u_{n,m}, \\ \mathbf{h}_n &= \mathrm{GELU}(W_{\downarrow}\, \mathrm{LN}(\mathbf{g}_n) + \mathbf{b}_{\downarrow}), \\ \Delta P_n &= 0.1 \rho_n \,\mathrm{reshape}_{N_q \times d_p} \left[\tanh(W_{\uparrow}\mathbf{h}_n + \mathbf{b}_{\uparrow})\right]. \end{aligned} \tag{16}$$

Here $W_{\downarrow} \in \mathbb{R}^{8\times M}$, $W_{\uparrow} \in \mathbb{R}^{N_q d_p \times 8}$, and $\mathbf{h}_n \in \mathbb{R}^8$. The up-projection weight and bias start at zero. Batch dimensions are omitted. The residual modifies only the learned prompt tokens; the projected object-query token is concatenated with them. The Q-Former output is projected into OPT's embedding space, giving a prefix with $N_q + 1$ tokens. With $\mathrm{CFPrefix}$ denoting this complete member pathway, the group prefix is

$$\mathbf{p}_n = \frac{1}{|\mathcal{I}_n|} \sum_{(i,k)\in\mathcal{I}_n} \mathrm{CFPrefix}(\mathbf{q}_i^k; P_n') \in \mathbb{R}^{(N_q+1)\times d_{\mathrm{OPT}}}. \tag{17}$$

The same fusion mapping is used in training and inference; caption targets and weak labels are used only for training. For the full CEG configuration, inference applies the residual directly, without the separate experimental sample router. Non-finite inputs, near-zero query norms, or reliability sums/maxima at most 10^{-8} trigger an inference fallback without a residual; training rejects such samples. The logarithm in Equation 12 uses the dtype's machine epsilon, a distinct safeguard.

Weak-label loss and reduction. Let N_{tr} be the number of caption-eligible training identities and $y_{n,m} \in \{0, 1\}$ the parsed target. Counts and positive-class weights are computed on those training identities:

$$N_m^+ = \sum_{n=1}^{N_{\mathrm{tr}}} y_{n,m}, \qquad \omega_m = \mathrm{clip}\left(\frac{N_{\mathrm{tr}} - N_m^+}{\max(N_m^+, 1)}, 1, 10\right). \tag{18}$$

Writing $z_{n,m} = \sum_{(i,k)\in\mathcal{I}_n} w_{i,k,m} a^k_{i,m}$ for the fused logits before admission, the per-identity loss is

$$\mathcal{L}_{\text{attr}}(n) = -\frac{1}{M}\sum_{m=1}^{M}\left[\omega_m y_{n,m}\log\sigma_{\text{sig}}(z_{n,m}) + (1-y_{n,m})\log(1-\sigma_{\text{sig}}(z_{n,m}))\right], \tag{19}$$

where σ_{sig} denotes sigmoid. The implementation uses numerically stable `BCEWithLogitsLoss` with `pos_weight` and the default mean reduction over the $1 \times M$ input. Positive-class weighting applies only to the positive term. Supervision acts on $z_{n,m}$, before $g_{n,m} = u_{n,m} z_{n,m}$; group labels supervise the fused representation, without separate per-view attribute annotations.

The parser encodes mentioned atoms as one and all remaining atoms as zero. These zeros represent missing caption mentions as well as genuine negatives. Accordingly, the weak-label probabilities characterize caption-derived evidence, and a confident zero is not an independently verified visual absence.

Let $\mathcal{U}_n$ index non-ignored, shifted caption-token targets $y^{\text{cap}}_{n,t}$. Teacher forcing and prefix anchoring use

$$\begin{aligned}\mathcal{L}_{\text{cap}}(n) &= -\frac{1}{|\mathcal{U}_n|}\sum_{t\in\mathcal{U}_n}\log p_\theta(y^{\text{cap}}_{n,t} \mid y^{\text{cap}}_{n,<t}, \mathbf{p}_n),\\ \mathcal{L}_{\text{prefix}}(n) &= \frac{\|\mathbf{p}_n - \mathbf{p}_n^{(0)}\|_F^2}{(N_q+1)d_{\text{OPT}}}.\end{aligned} \tag{20}$$

Here θ denotes the trainable caption pathway and CEG parameters. The frozen OPT decoder propagates gradients to the prefix; padding and prefix positions are ignored as targets. The cached initial prefix $\mathbf{p}_n^{(0)}$ is detached and stored in half precision. The prefix loss uses elementwise mean reduction with weight 1.0; the weak-label term has weight 0.05. One training identity contributes one optimization step.

Parameter accounting. The existing trainable caption pathway and the newly added CEG parameters are different counts. The former totals approximately 107.36M: a 33,600-parameter query adapter, approximately 105.16M Q-Former and learned-query-token parameters, and a 2,166,016-parameter language projection. CEG adds 245,200 parameters for evidence prediction and the low-rank prompt adapter. Thus the combined trainable count is approximately 107.61M. Frozen visual encoding, tracking, Re-ID, and OPT parameters are excluded from both trainable counts.

A.5 View-Count Diagnostics

This analysis stratifies the VECA+CEG and VECA+MVMean configurations in Table 3 by reference-view count. With VECA groups fixed, the in-domain ΔCIDEr is 0.89 for one-view identities and 55.03/38.48 for two-/three-view identities (Table 6).

Table 6: CEG–MVMean gain by reference-view count with VECA groups fixed. Metric deltas are shown on the $\times 100$ scale.

Split	Views	ΔCV-CHOTA	ΔGlobal-CapA	ΔCIDEr	ΔMETEOR	ΔSPICE
In-domain	1	+0.11	+0.29	+0.89	−0.02	+0.00
In-domain	2	+6.55	+20.06	+55.03	+4.05	+1.10
In-domain	3	+4.11	+14.17	+38.48	+2.76	+1.26
Cross-domain	3	+2.22	+12.87	+35.21	+2.78	+0.62
Cross-domain	4	+1.88	+10.75	+29.13	+1.96	+1.15

The one-view in-domain group changes little, whereas both multi-view in-domain groups improve in all three core caption-aware measures. Both cross-domain strata improve across all five reported metrics: the three- and four-view groups gain 35.21/29.13 CIDEr, 2.22/1.88 CV-CHOTA, and 12.87/10.75 Global-CapA, respectively. The two-view in-domain group has the largest CIDEr gain; the cross-domain gains in these three measures remain below those of the multi-view in-domain groups.

A.6 Evaluation Units and Metric Definitions

Result records link the configuration, seed, checkpoint digest, local-tracklet predictions, global identity groups, caption predictions, metric-input directory, and evaluator output. Component comparisons reuse the corresponding frozen tracklet records. Caption checkpoints are recorded per configuration, distinguishing separately trained models from inference-only interventions.

Caption evaluation uses one caption-eligible global identity as the scoring unit, irrespective of its number of frames or views. Missing reference captions remain null, and unaligned eligible references receive empty predictions. Invalid runs and unavailable source frames remain documented so that reconstruction preserves the evaluated population.

For scene-level identity assignment, IDTP, IDFP, and IDFN count identity true positives, false positives, and false negatives:

$$\mathrm{CVIDF1} = \frac{2\,\mathrm{IDTP}}{2\,\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}}. \tag{21}$$

At intersection-over-union threshold α, let $\mathcal{T}_\alpha$ contain matched detection pairs, and let $\mathcal{F}_\alpha$ and $\mathcal{N}_\alpha$ contain unmatched predicted and reference detections. For a match c, $\mathrm{TPA}_\alpha(c)$, $\mathrm{FPA}_\alpha(c)$, and $\mathrm{FNA}_\alpha(c)$ count true-positive, false-positive, and false-negative associations, respectively:

$$\begin{aligned} \text{CV-AssA}_\alpha &= \frac{1}{|\mathcal{T}_\alpha|} \sum_{c \in \mathcal{T}_\alpha} A_\alpha(c), \\ A_\alpha(c) &= \frac{\mathrm{TPA}_\alpha(c)}{\mathrm{TPA}_\alpha(c) + \mathrm{FPA}_\alpha(c) + \mathrm{FNA}_\alpha(c)}. \end{aligned} \tag{22}$$

Detection accuracy is

$$\mathrm{DetA}_\alpha = \frac{|\mathcal{T}_\alpha|}{|\mathcal{T}_\alpha| + |\mathcal{F}_\alpha| + |\mathcal{N}_\alpha|}. \tag{23}$$

HOTA averages over $\mathcal{A} = \{0.05, 0.10, \ldots, 0.95\}$:

$$\text{CV-HOTA} = \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \sqrt{\mathrm{DetA}_\alpha \, \text{CV-AssA}_\alpha}. \tag{24}$$