Context-Interactive Reasoning for Group Activity Detection

Xi Ai[1], Weihong Ren[1⋆], Shuhuan Han[1], Haoran Xu[1], Qian Dong[1], Pengyang Su[1], Zijian Wang[1], Shengchun Lin[1], Zhiyong Wang[1], and Honghai Liu[1,2]

[1] Harbin Institute of Technology, Shenzhen, China
25s153075@stu.hit.edu.cn, weihongren@hit.edu.cn
[2] Southeast University, China

Abstract. Group Activity Detection (GAD) aims to jointly infer group memberships and collective activities from videos. Existing approaches typically rely on actor-centric features to model group activities, assuming that collective semantics can be inferred solely from inter-actor relations. This overlooks critical context-interactive information, such as explicit spatial constraints and actor–scene dependencies, which are essential in crowded and multi-view environments. To address these limitations, we propose a Context-Interactive Reasoning framework that jointly models spatial topology and group semantic context. Specifically, a **S**hort-**T**erm **I**nteraction **R**egularizer (STIR) softly constrains inter-actor spatial relations with learnable frame-level topology priors, suppressing spurious connections and promoting spatially coherent grouping. Complementarily, a **L**ong-**R**ange **C**ontext **C**onditioner (LRCC) selectively incorporates global scene semantics via data-dependent gating, enabling activity-aware context utilization while avoiding unnecessary noise. Extensive experiments on challenging benchmarks demonstrate that our method achieves state-of-the-art performance in both group membership identification and collective activity recognition.

Keywords: Group Activity Detection · Social Interaction Understanding · Spatio-temporal Reasoning · Scene Context Modeling

1 Introduction

Group Activity Detection (GAD) aims to jointly infer group memberships and collective activities from videos [7,12,22,35], enabling a structured understanding of *who is with whom* and *what they are doing together*. Unlike conventional Group Activity Recognition (GAR), which typically assumes a single group per clip and pre-specified participants [8,17,18,34], GAD targets realistic scenes with multiple co-existing groups and outliers, where both grouping and activity recognition must be solved simultaneously. This capability is central to applications such as visual surveillance and social scene understanding [3, 26].

⋆ Corresponding author.

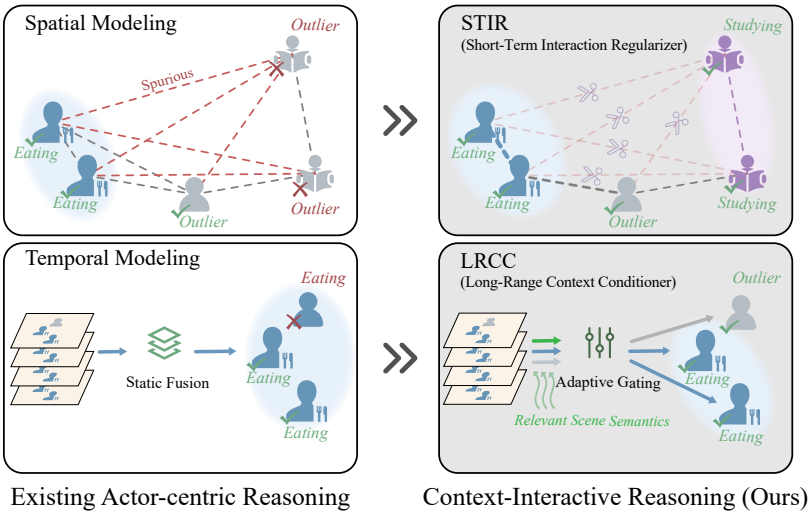

Fig. 1: Motivation for context-interactive reasoning. (*Left*) In crowded and occluded scenes, purely actor-centric reasoning can produce spatially implausible cross-group associations that blur group boundaries, while indiscriminate static fusion of scene context may inject background noise when global semantics are weakly relevant. (*Right*) We address these failure modes with a context-interactive design: **STIR** regularizes short-term inter-actor reasoning with topology-aware spatial priors to suppress implausible links, and **LRCC** adaptively injects clip-level scene semantics into actor/group tokens via data-dependent gating, enabling activity-aware context use without over-injection.

Recent progress in GAD has been accelerated by practical benchmarks, most notably the multi-camera Café dataset [22]. Café features dense crowds, frequent occlusions, diverse viewpoints, and a large proportion of outliers, requiring models to reason about semantic relations alongside spatial proximity. However, existing approaches still exhibit systematic limitations in such crowded, multi-view environments.

A major bottleneck is the prevailing actor-centric design: many methods model collective semantics primarily from inter-actor relations learned implicitly by attention over appearance features [14, 20, 24, 40]. In dense crowds, this often leads to spurious associations that violate the physical topology of social interactions [10,33], blurring group boundaries. While prior work sometimes imposes hard distance masks to encourage locality, such fixed thresholds can be brittle and may even suppress valid within-group interactions when the threshold is too small [11, 40, 46]. A second limitation is suboptimal use of global scene semantics. Ambiguous activities, such as eating versus studying, can be environment-conditioned and require actor-scene dependencies [3, 9, 31, 39, 45]; yet context is frequently ignored or fused statically, introducing background noise

when global scene semantics are irrelevant [26]. In our terminology, *short-term* and *long-range* describe the semantic scope of the two modules rather than the density of temporal sampling: STIR regularizes actor topology independently within each selected frame, whereas LRCC summarizes denser clip coverage into a clip-level scene-semantic prior.

Finally, stronger visual representations alone do not necessarily resolve group-level reasoning failures: recent GAD results suggest that gains from VFM backbones remain limited without task-aware structured reasoning on top [30]. Fig. 1 (*left*) illustrates two common failure modes of actor-centric reasoning in crowded scenes: spatially implausible cross-group associations under occlusion and indiscriminate static context fusion that introduces background noise. Fig. 1 (*right*) summarizes our remedy: STIR regularizes short-term spatial topology, while LRCC adaptively injects global scene semantics for activity-aware reasoning.

To address these limitations, we propose a Context-Interactive Reasoning framework that jointly models spatial topology and group semantic context for GAD. Specifically, a Short-Term Interaction Regularizer (STIR) injects explicit spatial constraints as a soft inductive bias into interaction reasoning, suppressing spurious connections and promoting spatially coherent grouping in dense scenes. Complementarily, a Long-Range Context Conditioner (LRCC) selectively incorporates global scene semantics via data-dependent gating, enabling activity-aware context utilization while avoiding unnecessary noise. Together, STIR and LRCC instantiate a unified context-interactive paradigm that resolves both structural ambiguity in grouping and semantic ambiguity in activity understanding. Our contributions are summarized as follows:

- We propose a **Context-Interactive Reasoning** framework for practical GAD that combines topology-aware interaction regularization with actor–scene semantic conditioning, targeting crowded and multi-view scenarios.
- We introduce **STIR**, a short-term regularizer that enforces topology-aware consistency over inter-actor relations through soft, explicit spatial constraints, effectively reducing spurious associations in dense crowds.
- We design **LRCC**, a long-range context conditioner that selectively injects global scene semantics via data-dependent gating, improving environment-conditioned activity recognition without introducing unnecessary noise.
- Extensive experiments on challenging benchmarks demonstrate state-of-the-art performance on both group membership identification and collective activity recognition.

2 Related Work

2.1 Group Activity Understanding: From GAR to Practical GAD

Early group activity understanding predominantly focused on GAR, where a single group is present per clip and participants are given, simplifying both grouping and activity inference [8, 18, 23, 25, 32]. To move toward realistic scenarios with multiple groups and outliers, GAD has emerged as a stricter formulation

that requires simultaneously localizing groups and recognizing their collective activities [13, 22, 35]. Recent GAR studies have also explored foundation-model adaptation with prompting [19], but they still operate under the conventional GAR setting rather than practical multi-group GAD.

Recent benchmarks such as Social-CAD [12] and JRDB-Act [13] extend earlier datasets with group annotations, but often contain many singleton groups, limiting their ability to stress-test multi-group reasoning [22]. Café was introduced to explicitly target practical GAD, providing multi-camera views, dense crowds, and rich annotations, and has become a central testbed for robust group reasoning [22].

Recent progress in practical GAD is driven by DETR-style set prediction frameworks that jointly localize groups and recognize activities in crowded multi-view scenes [5, 22, 47, 48], and is further advanced by richer interaction cues such as part-aware bottom-up group reasoning [21] as well as multimodal context integration via visual-language collaboration [26] and prompt-guided relational reasoning on top of vision foundation models [30].

2.2 Structured Interaction Reasoning with Spatial Priors

A core challenge in GAD is interaction reasoning under clutter and occlusion. Prior work explores structured relational modeling using graph neural networks and attention-based transformers [40, 42, 43].

Spatial cues are commonly incorporated in three ways: (i) as input features (*e.g.*, relative position/scale concatenated to actor tokens) learned by relation networks [1, 40]; (ii) as learned attention biases that modulate attention scores using spatial embeddings [14, 24]; or (iii) as hard adjacency constraints (*e.g.*, fixed distance masks/thresholds) to enforce locality [22,40], which are often non-adaptive across scenes, sensitive to threshold choice, and do not learn interaction ranges directly from data.

While hard masks can reduce implausible links between spatially distant actors, they may suppress valid interactions and generalize poorly across scenes with different crowd densities. These limitations motivate soft, interpretable, and learnable topology constraints that regularize relation modeling without relying on brittle heuristics. Recent practical GAD models on Café further highlight the importance of structured interaction reasoning under dense crowds [22]. Our STIR follows this line but differs by introducing a learnable head-wise spatial bias together with a soft, learnable distance penalty directly on attention logits, avoiding brittle hard thresholds while remaining fully differentiable.

2.3 Context Modeling for Collective Activities

Beyond grouping, collective activity recognition often depends on scene semantics and actor–scene dependencies [3, 31, 39, 45]. A common strategy is to aggregate global features and fuse them with actor tokens via concatenation or pooling, but such static fusion can be noise-prone when context is irrelevant [24, 41]. Recent work with VFMs further highlights that strong appearance features alone

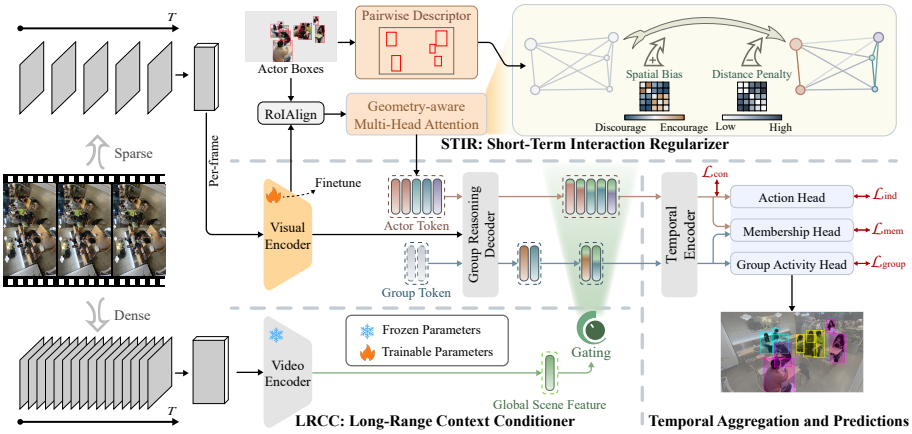

Fig. 2: Overall architecture of the proposed framework. A dual-stream representation extracts actor tokens from sparsely sampled frames and clip-level scene semantics from denser clip coverage. Both streams are uniformly sampled from the same clip segment; because LRCC consumes a pooled clip-level feature, no frame-wise alignment between the two streams is required. **STIR** regularizes short-term inter-actor relations before group reasoning, and **LRCC** adaptively conditions actor/group tokens on global scene semantics before temporal aggregation and prediction.

are insufficient: effective GAD requires structured reasoning mechanisms that explicitly model actor–group associations and contextual dependencies [30]. This calls for adaptive context utilization that can dynamically regulate when and how global semantics should influence local interaction reasoning. LRCC instantiates this idea with token-type-specific gating, allowing actor and group tokens to modulate global scene semantics differently under a shared clip-level semantic prior.

2.4 Vision Foundation Models for GAD

Self-supervised VFMs such as DINOv2 [28] and VideoMAE [36] provide transferable image and video representations. VLM-augmented dynamic group detection (DGD) further shows that foundation-model features can support group discovery when combined with temporal graph clustering [44]. Unlike DGD, practical GAD further requires group activity prediction and explicit outlier handling. More generally, generic pretraining does not explicitly capture the fine-grained actor relations required by GAD [6, 30], motivating task-specific structured reasoning on top of these features.

3 Method

3.1 Overview

Given a video clip $V = \{I_t\}_{t=1}^{T}$ with per-frame actor bounding boxes $\mathcal{B}_t = \{b_{t,i}\}_{i=1}^{N_t}$, Group Activity Detection (GAD) aims to predict a set of groups with memberships and collective activities, *i.e.*, $\{(\mathcal{G}_k, y_k)\}_{k=1}^{K}$, where $\mathcal{G}_k$ denotes the member set and y_k the group activity label, optionally accompanied by individual actions.

Our framework builds upon the set-prediction backbone of Practical GAD [22], where a fixed set of learnable group queries are matched to ground-truth groups via bipartite (Hungarian) matching. As shown in Fig. 2, we extract actor tokens on sparsely sampled frames using RoIAlign [15] and apply **STIR** to regularize short-term inter-actor relations with soft spatial constraints and suppress spatially implausible actor links. Next, we decode actor/group representations with the Practical GAD group decoder, and inject clip-level scene semantics computed from denser clip coverage into both actor and group tokens via **LRCC** using data-dependent gating to model actor–scene dependencies without introducing unnecessary noise. We further aggregate cross-frame evidence with our temporal encoder to stabilize grouping and activity prediction under occlusion and viewpoint changes, and finally use lightweight heads to predict memberships, collective activities, and individual actions. Training follows the same multi-task objective as Practical GAD.

3.2 Dual-Stream Visual Representation

Local actor tokens. For each frame I_t, we extract local visual features using a DINOv2 visual encoder, obtaining a feature map $\mathbf{F}_t$. Given actor boxes $\mathcal{B}_t$ in normalized image coordinates, we scale them to the backbone feature-map resolution and apply RoIAlign to extract actor-aligned features, yielding actor tokens $\mathbf{A}_t = \{\mathbf{a}_{t,i} \in \mathbb{R}^d\}_{i=1}^{N_t}$. In implementation, N_t is padded to a fixed upper bound, and invalid entries are masked by a validity mask. To encode coarse spatial layout, we add a learnable box positional embedding produced by a multilayer perceptron (MLP) over the normalized box coordinates. STIR is applied independently to each selected local frame; here, *short-term* refers to frame-level interaction scope rather than temporal sampling density.

Backbone adaptation strategy. To better adapt high-level representations to crowded interaction patterns while preserving the general semantics learned by pretraining, we fine-tune only the last few transformer blocks of the local DINOv2 encoder using a smaller learning rate than the task-specific modules. Key hyperparameters are reported in Sec. 4, with full details deferred to the supplementary material.

Global scene semantics. In parallel, we use a VideoMAE-v2 video encoder to obtain clip-level features $\mathbf{f}_{\text{mae}}$ that capture global spatiotemporal scene semantics. These global features are pre-extracted once per clip, kept fixed, and reused throughout GAD training and evaluation as the semantic source for LRCC. The

global stream uniformly samples a denser set of frames from the same clip segment and aggregates them into one clip-level feature. Since LRCC consumes this pooled representation rather than frame-wise features, no one-to-one temporal alignment with the local stream is required.

3.3 Short-Term Interaction Regularizer (STIR)

Crowded scenes often contain many visually similar actors, making implicit attention prone to forming spatially implausible cross-group relations that violate plausible image-plane spatial topology of social interactions. To mitigate this, STIR performs topology-aware regularization at the frame level by injecting a soft, learnable spatial topology prior into inter-actor relation modeling, suppressing implausible connections while preserving flexibility across different crowd densities and viewpoints.

Pairwise spatial descriptor. For a fixed frame t, we omit the frame index and denote boxes as $b_i = b_{t,i}$ and $b_j = b_{t,j}$ with $b_i = (x_i, y_i, w_i, h_i)$ and $b_j = (x_j, y_j, w_j, h_j)$ in normalized image coordinates. We define the spatial descriptor:

$$\mathbf{s}_{ij} = \left[\Delta x_{ij}, \Delta y_{ij}, \log \frac{w_i}{w_j}, \log \frac{h_i}{h_j}, \mathrm{dist}_{ij} \right], \tag{1}$$

where $(\Delta x_{ij}, \Delta y_{ij})$ are center offsets normalized by the image width and height respectively, and $\mathrm{dist}_{ij} = \sqrt{\Delta x_{ij}^2 + \Delta y_{ij}^2}$ denotes the pairwise center distance. This compact descriptor captures both relative displacement and relative scale, providing explicit spatial cues for interaction regularization.

Spatially Constrained Attention. STIR augments dot-product attention with two spatial terms: (i) a learnable spatial bias $\phi^h(\mathbf{s}_{ij})$ (implemented as an MLP that outputs per-head biases), and (ii) a distance-based soft suppression term that attenuates relations over large image-plane distances. For head h, the attention logit is:

$$\mathrm{score}_{ij}^h = \frac{(W_q^h \mathbf{a}_i)(W_k^h \mathbf{a}_j)^\top}{\sqrt{d_k}} + \phi^h(\mathbf{s}_{ij}) - \frac{\mathrm{dist}_{ij}^2}{\tau^2}, \tag{2}$$

where the per-head key dimension is $d_k = 64$. We parameterize $\tau = \exp(\rho)$ to ensure positivity and initialize τ to 1.0; it controls the effective image-plane interaction range. Compared with a hard distance threshold, this soft penalty is fully differentiable and learns the suppression range from data.

Masking and aggregation. To handle padded actors, we use a pairwise mask M_{ij} where $M_{ij} = 1$ indicates an invalid (padded) pair, and mask logits before softmax:

$$\mathrm{score}_{ij}^h \leftarrow -\infty \quad \text{if } M_{ij} = 1. \tag{3}$$

In practice, we implement the mask with a sufficiently large negative constant before softmax for numerical stability. The attention weights and per-head aggregation are then computed as

$$\mathrm{attn}_{ij}^h = \mathrm{softmax}_j(\mathrm{score}_{ij}^h), \qquad \mathbf{z}_i^h = \sum_j \mathrm{attn}_{ij}^h \cdot (W_v^h \mathbf{a}_j). \tag{4}$$

Finally, head outputs are concatenated and projected, followed by residual connections, layer normalization, and a feed-forward block:

$$\begin{aligned} \hat{\mathbf{a}}_i &= \mathrm{LN}\big(\mathbf{a}_i + W_o[\mathbf{z}_i^1; \cdots ; \mathbf{z}_i^H]\big)\,, \\ \mathbf{a}_i &\leftarrow \mathrm{LN}(\hat{\mathbf{a}}_i + \mathrm{FFN}(\hat{\mathbf{a}}_i))\,, \end{aligned} \tag{5}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization [2], $\mathrm{FFN}(\cdot)$ denotes a feed-forward network, and $\hat{\mathbf{a}}_i$ is an intermediate updated token. STIR is applied independently per frame with parameters shared across time, providing a stable short-term topology prior before group-level reasoning. By suppressing spatially implausible cross-group relations while retaining adaptable interaction ranges, STIR improves the spatial coherence of the inferred group structure in dense crowds.

3.4 Group Reasoning Decoder

After STIR refinement, we perform group reasoning using the group decoder of Practical GAD [22]. The decoder takes the refined per-frame actor tokens together with a fixed set of learnable group queries, and outputs decoded actor/group representations, denoted as $\{\mathbf{a}_{t,i}\}$ and $\{\mathbf{g}_{t,k}\}$. It alternates self-attention over the concatenated actor and group tokens and cross-attention to per-frame visual features, enabling group–actor interactions for joint grouping and activity reasoning. We keep this decoder unchanged to isolate the contributions of STIR and LRCC; the decoded tokens are then fed into LRCC and temporal modeling for clip-level semantic conditioning and cross-frame aggregation.

3.5 Long-Range Context Conditioner (LRCC)

Recognizing collective activities often depends on actor-scene dependencies (*e.g.*, eating vs. studying), where local appearance and interactions alone may be ambiguous. However, naively fusing global scene semantics can be noise-prone when scene cues are weakly relevant. To this end, LRCC selectively injects macro-scale scene semantics via data-dependent gating, in the spirit of conditional feature modulation [29], enabling activity-aware context conditioning while avoiding indiscriminate context over-injection.

Projection to token space. Given a clip-level global feature $\mathbf{f}_{\mathrm{mae}}$, we first map it into the same embedding space as actor/group tokens:

$$\mathbf{z}_{\mathrm{global}} = \mathrm{Proj}(\mathbf{f}_{\mathrm{mae}}) \in \mathbb{R}^d, \tag{6}$$

where $\mathrm{Proj}(\cdot)$ maps the VideoMAE-v2 feature to the actor/group token dimension $d = 256$ using a linear layer followed by LayerNorm, ReLU activation, and dropout.

Actor/group-specific gating. Let $\mathbf{a}_i$ denote an actor token and $\mathbf{g}_k$ denote a group token. LRCC computes two non-shared gates to condition the two token types differently:

$$\begin{aligned} \alpha_i &= \sigma\big(W_a[\mathbf{a}_i; \mathbf{z}_{\text{global}}]\big), & \tilde{\mathbf{a}}_i &= \mathbf{a}_i + \alpha_i \, \mathbf{z}_{\text{global}}; \\ \beta_k &= \sigma\big(W_g[\mathbf{g}_k; \mathbf{z}_{\text{global}}]\big), & \tilde{\mathbf{g}}_k &= \mathbf{g}_k + \beta_k \, \mathbf{z}_{\text{global}}, \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, $[\cdot;\cdot]$ denotes concatenation, and $\alpha_i, \beta_k \in (0,1)$ are per-token scalar gates. We adopt scalar gates for lightweight and stable conditioning, which reduce parameter overhead, improve interpretability of token-wise context injection, and avoid over-parameterized channel-wise modulation under limited supervision. This asymmetric design reflects the different roles of actor tokens (instance-level cues) and group tokens (group-level semantics) in downstream reasoning, and allows LRCC to regulate context influence in a token-dependent manner.

Placement and temporal sharing. In the full pipeline, LRCC is applied after group reasoning and before temporal modeling. Since $\mathbf{f}_{\text{mae}}$ is clip-level, the projected context $\mathbf{z}_{\text{global}}$ is shared across frames within the same clip and broadcast to all actor/group tokens before gating. This yields a stable long-range semantic prior while preserving adaptive, token-wise injection strength.

3.6 Temporal Modeling

To capture cross-frame dynamics, we use a temporal encoder composed of a residual temporal convolution block followed by stacked multi-head temporal self-attention layers. Given a token trajectory $\mathbf{U} = [\mathbf{u}_1; \ldots; \mathbf{u}_T] \in \mathbb{R}^{T \times d}$, we first apply a residual temporal convolution:

$$\hat{\mathbf{U}} = \mathbf{U} + \text{TCN}(\mathbf{U}), \tag{8}$$

where $\text{TCN}(\cdot)$ is a lightweight residual 1D temporal convolution block [4] with normalization, nonlinearity, and dropout. We then model long-range temporal dependencies with multi-head scaled dot-product self-attention [37]:

$$\text{TSA}(\hat{\mathbf{U}}) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{9}$$

where Q, K, V are linear projections of the input sequence, and $d_k = 64$ is the per-head key dimension. For brevity, each temporal layer $\text{TSA}^{(l)}(\cdot)$ denotes a complete temporal Transformer block (self-attention + FFN with residual connections and layer normalization; head index omitted for clarity).

We stack L temporal layers:

$$\mathbf{U}^{(0)} = \hat{\mathbf{U}} + \mathbf{P}, \qquad \mathbf{U}^{(l+1)} = \text{TSA}^{(l)}(\mathbf{U}^{(l)}),\ l = 0, \ldots, L-1, \tag{10}$$

where $\mathbf{P} \in \mathbb{R}^{T \times d}$ is a learnable temporal positional embedding. After temporal encoding, we obtain a clip-level representation via lightweight attention pooling:

$$\mathbf{u}_{\text{clip}} = \sum_{t=1}^{T} \omega_t \, \mathbf{u}_t^{(L)}, \qquad \omega_t = \text{softmax}_t(w^\top \mathbf{u}_t^{(L)}), \tag{11}$$

and we instantiate $\mathbf{u}_t$ as $\tilde{\mathbf{a}}_{t,i}$ for actor tokens or $\tilde{\mathbf{g}}_{t,k}$ for group tokens.

3.7 Prediction and Training

Following Practical GAD [22], we use lightweight feed-forward heads for individual actions and group activities, and a dot-product membership head for actor–group assignment. During training, predicted group slots are matched to ground-truth groups via bipartite matching, and unmatched slots are supervised as no-group.

We adopt the same multi-task objective as Practical GAD [22], including $\mathcal{L}_{\text{ind}}$, $\mathcal{L}_{\text{group}}$, $\mathcal{L}_{\text{mem}}$, and an InfoNCE-style [27] consistency regularizer $\mathcal{L}_{\text{con}}$. Specifically, $\mathcal{L}_{\text{con}}$ is a temperature-scaled contrastive loss on actor embeddings, where actors belonging to the same group are treated as positive pairs and actors from different groups are treated as negative pairs, encouraging intra-group compactness and inter-group separation. The overall objective is:

$$\mathcal{L} = \lambda_{\text{ind}}\mathcal{L}_{\text{ind}} + \lambda_{\text{group}}\mathcal{L}_{\text{group}} + \lambda_{\text{mem}}\mathcal{L}_{\text{mem}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{12}$$

4 Experiments

4.1 Datasets and Evaluation Protocol

Café. We primarily evaluate on the Café benchmark [22], which targets practical GAD with dense crowds, frequent occlusions, multi-camera views, and a large proportion of outliers. Following the official protocol, we report results under both split-by-view and split-by-place to assess cross-view and cross-scene generalization. All numbers are obtained using the official evaluation code and the same data splits as prior work for fair comparison.

Social-CAD. We additionally evaluate on Social-CAD [12] as a complementary benchmark with less crowded scenes. While Social-CAD is less challenging than Café in terms of crowd density and multi-view ambiguity, it provides a useful sanity check for verifying that the proposed reasoning components do not overfit to a single benchmark.

Metrics. For Café, we follow [22] and report Group mAP at two IoU thresholds ($\text{mAP}_{1.0}$ / $\text{mAP}_{0.5}$) and Outlier mIoU, which jointly measure group localization/assignment quality and outlier handling. For Social-CAD, following prior work, we report social activity accuracy and individual action accuracy under the benchmark protocol.

4.2 Implementation Details

Our model takes as input a video clip with per-frame actor bounding boxes. Following the official protocol of [22], we use ground-truth actor boxes for all experiments. We use a partially fine-tuned DINOv2-Base visual encoder for local actor features, with a reduced learning rate for the unfrozen blocks and the remaining blocks kept frozen. In contrast, the VideoMAE-v2 [38] video encoder is kept fully frozen and used to pre-extract clip-level features offline as the global semantic source for LRCC. All reasoning modules are trained end-to-end on top

of these features. We otherwise follow the training recipe and optimization strategy of Practical GAD [22]; key settings are summarized below, with remaining details provided in the supplementary material.

Key settings. Specifically, we use DINOv2-Base as the local visual encoder and fine-tune its last 2 transformer blocks with a learning rate set to $0.1\times$ that of the task-specific modules. For global scene semantics, we use VideoMAE-v2 Giant and keep it fully frozen, extracting clip-level features offline. For temporal sampling, both streams are uniformly sampled from the same clip segment. On Café, the local and global streams use 8 and 16 frames, respectively; on Social-CAD, they use 1 and 16 frames, respectively. STIR is applied independently to each local frame, whereas VideoMAE-v2 aggregates the global frames into one clip-level feature, so no frame-wise alignment is required.

4.3 Comparison with State-of-the-Art on Café

Tab. 1 compares our approach with state-of-the-art methods on Café. Our method establishes the best results among the reported entries in our comparison table, with strong gains on both split-by-view and split-by-place.

We note that ProGraD reports frozen-backbone results on both splits and an additional fully fine-tuned result only on split-by-place; we list the two settings separately in Tab. 1. On split-by-view, our method improves the strongest prior result for each metric by +10.80, +11.08, and +5.56, respectively. On the more challenging split-by-place, our method consistently outperforms the fully fine-tuned ProGraD setting by +3.40, +3.83, and +4.26, respectively.

These consistent margins indicate that the improvement is not tied to a specific camera view or scene. We attribute the gains to the proposed context-interactive reasoning: STIR stabilizes group partitioning by suppressing spatially implausible cross-group associations in dense crowds, while LRCC improves environment-conditioned activity understanding by injecting scene semantics only when needed. Moreover, these improvements are achieved without modifying the group decoder or supervision protocol, indicating that STIR/LRCC can serve as plug-and-play reasoning modules on top of an established GAD backbone.

4.4 Results on Social-CAD

Tab. 2 reports results on Social-CAD. Despite being less crowded than Café, our method remains strong. Notably, on Social-CAD we sample only one local frame for STIR while still using 16-frame clip coverage for LRCC, achieving 72.8% social accuracy and 90.2% individual accuracy. This surpasses the best prior social accuracy (70.0% of ProGraD) and improves over prior methods that report individual action accuracy, suggesting that the proposed interaction regularization and selective context conditioning generalize beyond Café-specific biases. We view Social-CAD as a complementary benchmark with reduced crowd ambiguity, and the consistent gains suggest that our modules improve reasoning robustness rather than overfitting to Café-specific crowd patterns. We note that

Table 1: Comparison with state-of-the-art methods on the Café dataset. Methods use different visual encoders and training recipes; controlled backbone ablations are provided in Tab. 3. †ProGraD reports frozen-backbone results on both splits and an additional fully fine-tuned result only on split-by-place.

Method	Year	Split by view			Split by place		
		Group $mAP_{1.0}$	Group $mAP_{0.5}$	Outlier mIoU	Group $mAP_{1.0}$	Group $mAP_{0.5}$	Outlier mIoU
ARG [40]	2019	11.03	34.50	56.61	6.87	28.44	46.72
Joint [12]	2020	14.05	36.08	60.09	8.39	26.26	55.95
JRDB-base [13]	2022	15.43	34.81	60.43	9.42	25.75	48.00
HGC [35]	2022	6.55	26.29	56.84	3.47	18.46	52.56
Practical GAD [22]	2024	18.84	37.53	67.64	10.85	30.90	63.84
VLCMTN [26]	2025	19.67	41.06	68.29	11.58	30.98	65.48
ProGraD (frozen) [30]	2026	16.85	43.42	70.42	17.03	39.11	67.89
ProGraD (full FT)† [30]	2026	–	–	–	20.42	44.14	67.71
Ours	2026	**30.47**	**54.50**	**75.98**	**23.82**	**47.97**	**71.97**

Table 2: Quantitative results on Social-CAD. †Individual action accuracy was not reported for Practical GAD or ProGraD.

Method	Backbone	Social Accuracy	Individual Accuracy†
ARG [40]	Inception-v3	49.0	78.4
Joint [12]	I3D	69.0	83.3
Practical GAD [22]	ResNet-18	69.2	–
ProGraD [30]	DINOv2	70.0	–
STIRM [16]	ResNet-50	69.6	83.1
Ours	DINOv2	**72.8**	**90.2**

some prior works report only social activity accuracy on Social-CAD; our table follows the metrics available in the respective papers while evaluating our model under the standard benchmark protocol.

4.5 Ablation Studies

Unless otherwise stated, all ablations are conducted on the Café split-by-place setting, which is the more challenging split due to its stronger cross-scene generalization requirement. We keep the same temporal aggregation module in all ablations to stabilize predictions under occlusion and viewpoint changes and to ensure fair comparisons that isolate the contributions of STIR and LRCC. Additional ablations on temporal design choices and other implementation details are provided in the supplementary material.

Component contribution. To disentangle the contribution of each module, Tab. 3 reports controlled ablations with (i) our baseline built on the Practical

Table 3: Component contribution analysis. The configuration used in the final model is highlighted in gray.

Settings	Backbone	Components		Split by place		
		STIR	LRCC	Group $mAP_{1.0}$	Group $mAP_{0.5}$	Outlier mIoU
Baseline	ResNet-18	✗	✗	8.75	29.65	59.86
	DINOv2			10.45	31.76	65.75
+STIR Only	ResNet-18	✓	✗	15.86	36.71	68.44
	DINOv2			15.69	35.16	69.71
+LRCC Only	ResNet-18	✗	✓	12.71	34.83	66.97
	DINOv2			18.98	35.58	69.21
Full	ResNet-18	✓	✓	17.14	34.45	69.64
	DINOv2			**23.82**	**47.97**	**71.97**

GAD decoder and the same temporal aggregation module, (ii) +STIR only, (iii) +LRCC only, and (iv) the full model. We also evaluate each configuration with two appearance encoders (ResNet-18 and DINOv2-Base) to verify that gains are not merely due to stronger features.

Tab. 3 shows that STIR consistently strengthens grouping robustness by suppressing spatially implausible cross-group associations, while LRCC provides complementary gains by selectively injecting scene semantics when beneficial. Notably, simply upgrading the visual encoder does not reliably resolve group-level reasoning failures, whereas adding STIR/LRCC yields consistent improvements, with the clearest gains observed when paired with the stronger backbone. Overall, STIR and LRCC exhibit the clearest complementarity with DINOv2: STIR improves *who is with whom* via topology-aware regularization, while LRCC improves *what they are doing together* via selective semantic conditioning.

STIR variants. Tab. 4 compares alternative ways of injecting spatial priors into interaction reasoning. Among the evaluated variants, the full STIR design performs best, indicating that a soft, learnable topology constraint is more effective than hard masking or bias-only alternatives in dense crowds. Attention bias and hard mask both improve over w/o STIR, but they are respectively less constrained by explicit topology and more sensitive to threshold choice and scene density. These results support our design choice of combining learnable spatial bias with soft distance-based suppression.

LRCC variants. Tab. 5 evaluates LRCC against static fusion strategies. Here, "Global pooling" denotes a static baseline that broadcasts the projected clip-level context to all actor/group tokens and fuses it uniformly without token-dependent gating, *i.e.*, the same context contribution is applied to all tokens. Two-branch adaptive gating consistently outperforms both static fusion and a shared gate, supporting token-type-specific modulation of global scene semantics for actor and group tokens. Static fusion (concat/add/pooling) provides partial gains but

Table 4: Internal analysis of STIR (split-by-place). The configuration used in the final model is highlighted in gray.

Variant	Group $mAP_{1.0}$	Group $mAP_{0.5}$	Outlier mIoU
w/o STIR	18.98	35.58	69.21
Spatial bias	22.64	45.10	71.39
Hard mask	20.70	43.66	70.00
STIR	**23.82**	**47.97**	**71.97**

Table 5: Internal analysis of LRCC (split-by-place). The configuration used in the final model is highlighted in gray.

Variant	Group $mAP_{1.0}$	Group $mAP_{0.5}$	Outlier mIoU
w/o LRCC	15.69	35.16	69.71
Concat	17.57	35.81	70.36
Add	16.59	38.31	68.21
Global pooling	22.90	43.28	70.93
Shared gate	16.45	32.41	68.31
Two-branch gates	**23.82**	**47.97**	**71.97**

may inject noise when scene context is weakly relevant, while a shared gate is less flexible than the two-branch design.

4.6 Qualitative Results

Fig. 3 compares our method with Practical GAD in crowded scenes containing multiple spatially adjacent groups and outliers. Under substantial actor overlap and occlusion, the baseline frequently merges nearby but semantically distinct groups, assigns outliers to an incorrect group, or consequently predicts an incorrect group activity. For example, in the first and third rows, the baseline fails to preserve the distinct *Selfie* group located close to the *Fighting* group. In the second row, it produces less accurate assignments around the *Working* group and nearby outliers, while in the fourth row it fails to separate the neighboring *Eating* and *Fighting* groups. In contrast, our predictions more closely match the ground-truth group partitions, recover the corresponding activity labels, and keep unrelated outliers separated from valid groups.

These qualitative patterns are consistent with the intended roles of the two proposed modules: STIR reduces spatially implausible inter-actor associations, whereas LRCC selectively incorporates clip-level scene semantics when they are relevant to actor and group representations. Together, they improve both group-boundary preservation and activity prediction in crowded, multi-view environments.

5 Conclusion

In this paper, we study practical Group Activity Detection (GAD) in crowded multi-view scenes, where occlusion-induced spurious relations and noisy context distract actor-centric reasoning. We propose a Context-Interactive Reasoning framework on a set-prediction GAD backbone, combining topology-aware interaction regularization with adaptive scene-semantic conditioning. Concretely,

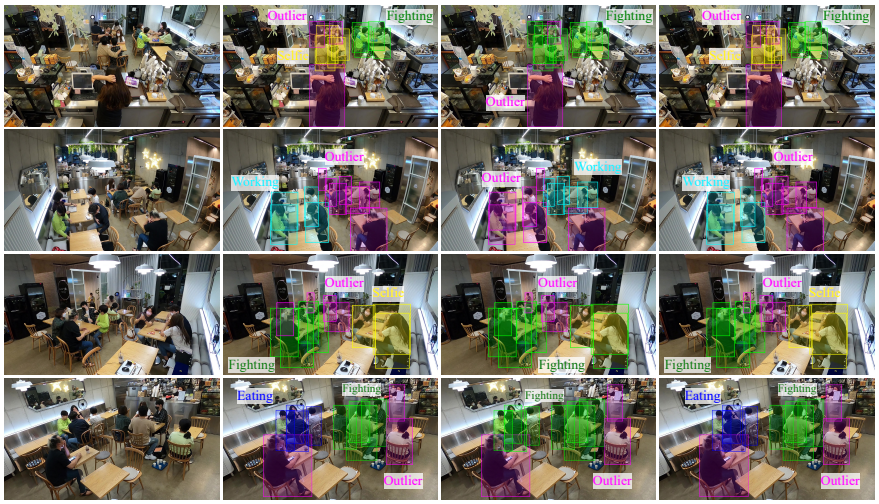

Fig. 3: Qualitative comparisons on Café under the split-by-place setting. Each row presents a representative test frame from a different crowded scene. Within each panel, boxes with the same color denote the same group, while the overlaid text indicates the corresponding group activity or outlier label. (a) Input frame, (b) Ground-truth, (c) Prediction of Practical GAD, and (d) Prediction of our model.

STIR injects soft spatial topology priors into inter-actor relation modeling to suppress spatially implausible links while preserving learnable interaction ranges. Complementarily, LRCC selectively injects clip-level scene semantics through token-dependent gating, enabling activity-aware context usage without indiscriminate context fusion. A temporal encoder further aggregates cross-frame evidence to stabilize grouping and activity prediction under viewpoint changes. Extensive experiments on challenging benchmarks demonstrate state-of-the-art performance for both group membership identification and collective activity recognition, and validate that structured context interaction remains crucial even with strong vision foundation model features.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2025YFE0218500, in part by the National Natural Science Foundation of China under Grants 62573163 and 62503139, in part by the Guangdong Basic and Applied Basic Research Foundation under Grants 2024A1515012028 and 2026A1515011822, and in part by the Shenzhen Science and Technology Program under Grants JCYJ20250604145514019 and GXWD20231129125006001.

References

1. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: CVPR. pp. 7892–7901 (2019)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: CVPR. pp. 4315–4324 (2017)
4. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: CVPR. pp. 14455–14465 (2024)
7. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Discovering groups of people in images. In: ECCV. pp. 417–433. Springer (2014)
8. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: ICCV Workshops. pp. 1282–1289. IEEE (2009)
9. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR. pp. 3273–3280. IEEE (2011)
10. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: BMVC. vol. 2, pp. 23.1–23.12 (2011)
11. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: CVPR. pp. 4772–4781 (2016)
12. Ehsanpour, M., Abedin, A., Saleh, F., Shi, J., Reid, I., Rezatofighi, H.: Joint learning of social groups, individuals action and sub-group activities in videos. In: ECCV. pp. 177–195. Springer (2020)
13. Ehsanpour, M., Saleh, F., Savarese, S., Reid, I., Rezatofighi, H.: JRDB-Act: A large-scale dataset for spatio-temporal action, social group and activity detection. In: CVPR. pp. 20983–20992 (2022)
14. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.: Actor-transformers for group activity recognition. In: CVPR. pp. 839–848 (2020)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV. pp. 2961–2969 (2017)
16. Huang, J., Li, L., Qing, L., Tang, W., Wang, P., Guo, L., Peng, Y.: Spatio-temporal interactive reasoning model for multi-group activity recognition. PR **159**, 111104 (2025)
17. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: ECCV. pp. 721–736 (2018)
18. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR. pp. 1971–1980 (2016)
19. Jin, Z., Feng, A., Chemburkar, A., De Melo, C.M.: Promptgar: Flexible promptive group activity recognition. In: WACV. pp. 4461–4471 (2026)

20. Karki, D., Ramazanova, M., Cioppa, A., Giancola, S., Ghanem, B.: Pixels or positions? benchmarking modalities in group activity recognition. In: CVPR Workshops. pp. 9998–10008 (2026)
21. Kim, D., Cho, M., Kwak, S.: Part-aware bottom-up group reasoning for fine-grained social interaction detection. In: NeurIPS. vol. 38, pp. 138384–138410 (2026)
22. Kim, D., Song, Y., Cho, M., Kwak, S.: Towards more practical group activity detection: A new benchmark and model. In: ECCV. pp. 240–258. Springer (2024)
23. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR. pp. 1354–1361. IEEE (2012)
24. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: ICCV. pp. 13668–13677 (2021)
25. Li, X., Choo Chuah, M.: Sbgar: Semantics based group activity recognition. In: ICCV. pp. 2876–2885 (2017)
26. Nian, F., Lu, W., Wang, J., Li, C., Fu, Y., Wu, Z.: Visual-language collaborative multimodal transformer network for group activity detection in surveillance videos. Multimedia Systems **31**(4), 305 (2025)
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
29. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI. vol. 32 (2018)
30. Ponbagavathi, T.T., Yang, C., Roitberg, A.: Structured relational reasoning for group activity assessment. In: CVPR Workshops. pp. 8931–8940 (2026)
31. Pramono, R.R.A., Chen, Y.T., Fang, W.H.: Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: ECCV. pp. 71–90. Springer (2020)
32. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: STAGNet: An attentive semantic rnn for group activity recognition. In: ECCV. pp. 101–117 (2018)
33. Setti, F., Russell, C., Bassetti, C., Cristani, M.: F-formation detection: Individuating free-standing conversational groups in images. PloS one **10**(5), e0123783 (2015)
34. Shu, T., Todorovic, S., Zhu, S.C.: CERN: confidence-energy recurrent network for group activity recognition. In: CVPR. pp. 5523–5531 (2017)
35. Tamura, M., Vishwakarma, R., Vennelakanti, R.: Hunting group clues with transformers for social group activity recognition. In: ECCV. pp. 19–35. Springer (2022)
36. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS **35**, 10078–10093 (2022)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
38. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling video masked autoencoders with dual masking. In: CVPR. pp. 14549–14560 (2023)
39. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: CVPR. pp. 3048–3056 (2017)
40. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: CVPR. pp. 9964–9974 (2019)

41. Yan, R., Tang, J., Shu, X., Li, Z., Tian, Q.: Participation-contributed temporal dynamic model for group activity recognition. In: ACM MM. pp. 1292–1300 (2018)
42. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Social adaptive module for weakly-supervised group activity recognition. In: ECCV. pp. 208–224. Springer (2020)
43. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: HiGCIN: Hierarchical graph-based cross inference network for group activity recognition. IEEE TPAMI **45**(6), 6955–6968 (2023)
44. Yokoyama, K., Nakatani, C., Ukita, N.: Dynamic group detection using VLM-augmented temporal groupness graph. In: ICCV. pp. 10475–10484 (October 2025)
45. Yuan, H., Ni, D.: Learning visual context for group activity recognition. In: AAAI. vol. 35, pp. 3261–3269 (2021)
46. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. NeurIPS **17** (2004)
47. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: CVPR. pp. 16965–16974 (2024)
48. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)