TEMPOGROUND: STATE-AWARE STREAMING VISUAL GROUNDING WITH VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding maps language referents to spatial targets and is central to open-vocabulary perception with vision-language models. Existing methods have made substantial progress on single-frame and video-based visual grounding, yet under streaming inputs they still suffer from identity drift, cross-frame inconsistency, and fragile localization under partial occlusion. To address these issues, we present **TempoGround**, a VLM-native framework that detects cross-frame object correspondence and explicitly models object presence states, thereby enabling accurate and consistent visual grounding under streaming inputs. The key is a curriculum prediction mechanism guided by state-aware cross-frame correspondence: TempoGround resolves 2D instance association, predicts whether each object newly enters, continues in, or leaves the view, decodes the 2D box, and then lifts it to a camera-frame 3D box. As token-level supervision alone cannot capture the geometric objectives of streaming grounding, we further introduce Streaming Grounding Reinforcement (SGR), which optimizes TempoGround with verifiable Grounding, Identity, and Consistency rewards, jointly reinforcing persistent localization and temporally consistent predictions. We carefully design a three-stage training strategy and train TempoGround on large-scale data. We evaluate visual grounding under causally streaming inputs on multiple challenging benchmarks: TempoGround improves $\text{F1}_{2D}@0.5$ and $\text{F1}_{2D}@0.95$ by 4.4 and 0.5 on average, and $\text{F1}_{3D}@0.25$ and AP_{3D} by 6.2 and 7.5, respectively. These results demonstrate that TempoGround provides a practical foundation for visual grounding under streaming inputs.

1 INTRODUCTION

Vision-language models (VLMs) have made visual grounding a practical route to open-vocabulary perception: given a free-form language query, the model localizes the referred objects as spatially actionable targets, which is important for embodied perception and related downstream control. Most existing VLM grounders operate on single-frame or video-based inputs. In practical deployments, however, sensory observations are delivered as a causal stream. We define streaming visual grounding as the setting in which a language query is given, images arrive incrementally, and the model must continuously track the matching objects in view and produce their 2D/3D bounding boxes. As shown in Figure 1, streaming visual grounding commonly exhibits three characteristic failure modes. Identity drift confuses same-category instances across time. Cross-frame inconsistency yields temporally unstable boxes across adjacent frames. Fragile localization under partial occlusion arises when intermittent visibility causes missed or degraded detections that cannot be corrected without causal cross-frame reasoning.Consequently, visual grounding under streaming inputs remains challenging.

Recent 3D scene understanding methods enhance object-level perception and grounding by injecting globally aligned geometric information. These approaches typically perform object grounding in a world or scene coordinate system and therefore introduce a global coordinate dependency. For object-level perception under ego-centric viewpoints, as in embodied manipulation and related scenarios, world-frame predictions are not directly applicable in the current camera view and require an additional alignment step that introduces additional alignment error. We therefore aim for a VLM-native method for streaming visual grounding that consumes causal image streams and predicts in the current camera frame.

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.

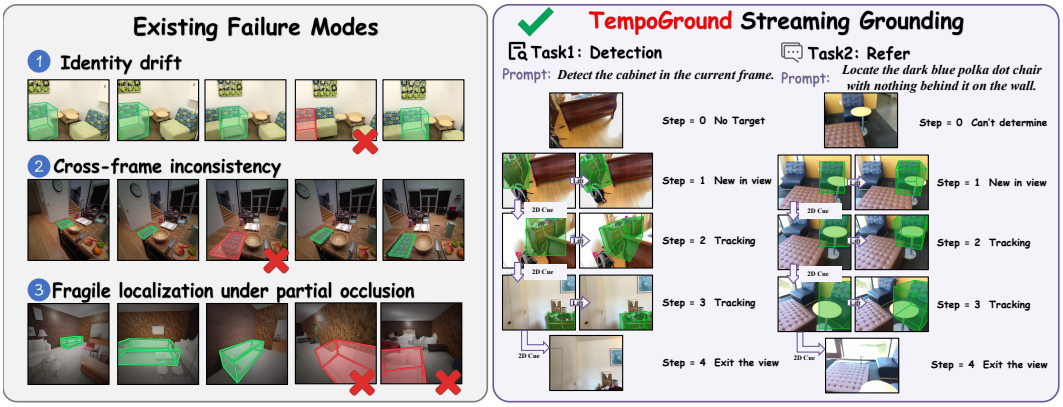

Figure 1: **TempoGround for streaming visual grounding.** Under streaming inputs, visual grounding commonly exhibits three characteristic failure modes (left). By exploiting cross-frame correspondence for state-aware curriculum prediction, TempoGround yields more stable camera-frame 2D/3D localization under streaming inputs (right) for both detection and referring grounding.

To address this limitation, we present TempoGround with a state-aware streaming curriculum prediction mechanism that leverages cross-frame correspondence for object state prediction and spatial localization. At each timestep, under language guidance, TempoGround first associates instances across adjacent frames by matching a lightweight 2D cue from the previous prediction against corresponding regions in the current image, thereby maintaining consistent object identity. Based on this correspondence, TempoGround then estimates whether each object newly enters, persists in, or exits the field of view. Conditioned on these correspondence and presence cues, TempoGround better captures cross-frame object states and spatial localization. As reliable 3D predictions benefit from stable 2D evidence Man et al. (2026), TempoGround first predicts a 2D box for each instance and then lifts it to a camera-frame 3D box. Taken together, TempoGround follows a curriculum prediction chain that resolves 2D correspondence, estimates object presence, decodes the 2D box, and lifts the prediction to 3D. This design allows TempoGround to exploit cross-frame instance correspondence and spatial position changes under streaming inputs, supporting more robust and accurate tracking.

As a VLM-native approach to streaming object grounding, TempoGround is built on a standard VLM decoder, produces outputs in a standard autoregressive manner, and is trained with standard supervised fine-tuning on carefully constructed large-scale data. However, supervised fine-tuning alone remains insufficient, because token-level supervision treats bounding box prediction as an autoregressive token sequence and therefore does not directly model the geometric attributes of visual bounding boxes, nor does it provide direct box-level supervision under streaming inputs. To meet the consistency requirements of streaming grounding, we introduce Streaming Grounding Reinforcement (SGR), an RLVR stage that optimizes TempoGround with verifiable Grounding, Identity, and Consistency rewards. These rewards target localization quality along the stream, correct enter and exit decisions, and stable cross-frame predictions.

Building on the above considerations, we construct a three-stage training recipe that progresses from single-frame geometric competence to streaming supervision and finally to sequence-level reinforcement. Stage 1 applies supervised fine-tuning on **26M** samples to strengthen large-scale 2D detection and 2D-to-3D lifting, establishing the spatial foundation for subsequent learning. Building on this competence, Stage 2 continues with supervised fine-tuning on **5.1M** samples to establish curriculum prediction, enabling cross-frame correspondence to support presence estimation and consistent localization. Stage 3 then refines these streaming capabilities with SGR under verifiable rewards, aligning training with the geometric objectives that token-level supervision alone cannot capture.

Our contributions are as follows:

- We propose **TempoGround**, a VLM-native framework that enables accurate and consistent streaming visual grounding through **curriculum prediction** guided by state-aware cross-

frame correspondence, from 2D association and presence estimation to camera-frame 3D grounding.

- We introduce **Streaming Grounding Reinforcement (SGR)**, an RLVR method that optimizes TempoGround with verifiable Grounding, Identity, and Consistency rewards to align training with the geometric objectives of streaming grounding.
- Extensive experiments on multiple challenging benchmarks under causally streaming inputs show that TempoGround improves $F1_{2D}$ at IoU thresholds 0.5 and 0.95 by 4.4 and 0.5 on average, and improves $F1_{3D}$ at IoU 0.25 and AP_{3D} by 6.2 and 7.5, providing a practical foundation for streaming visual grounding.

2 Related Works

2.1 VLM-based Visual Perception

Vision-language models have been increasingly applied to spatial understanding in 3D environments. Reconstructed geometry or multi-view video can be encoded into LLMs for relational and layout-level dialogue Hong et al. (2023), coupled with embodied reasoning and control Huang et al. (2023), or unified with grounding, captioning, and question answering under shared instruction following Huang et al. (2024); Zheng et al. (2025). Under camera motion, explicit geometry is further injected through position-aware visual tokens Zhu et al. (2024), globally consistent geometric fusion Zheng et al. (2026), or reconstructive 3D tokens for spatial reasoning Fan et al. (2025), while first-person 3D sensing resources continue to expand Wang et al. (2024). These methods improve scene-level semantics and spatial reasoning over offline reconstructions and short clips, yet their supervision and inference remain largely scan- or clip-centric, and object predictions are often formulated in a world or scene coordinate system. TempoGround instead targets object-level localization in the current camera frame under causally arriving image streams.

2.2 Visual Object Grounding

Visual object grounding associates language queries with spatial targets. Open-vocabulary detectors condition localization on category names or referring expressions Liu et al. (2024); Chen et al. (2020), while multimodal LLMs cast boxes as autoregressive tokens Peng et al. (2024); Bai et al. (2025); Chen et al. (2023). Subsequent work strengthens this generative formulation through pixel-aligned grounding Rasheed et al. (2024), high-throughput generative detection Jiang et al. (2025); Wang et al. (2026), and staged 2D-to-3D prediction in the camera frame Man et al. (2026). Reinforcement with verifiable rewards further aligns generated geometry with evaluation metrics Linghu et al. (2026). These advances mainly address single-frame or offline clip settings, where each observation can be treated largely independently or with access to the full sequence. Streaming visual grounding additionally requires identity-stable and cross-frame-consistent box updates as images arrive causally. TempoGround addresses this setting.

3 Method

We present TempoGround, a VLM-native framework for streaming visual grounding. Figure 2 overviews the state-aware curriculum prediction interface and the progressive three-stage training recipe. Section 3.1 formalizes the streaming visual grounding task and the geometry conventions used throughout. Section 3.2 describes curriculum prediction guided by state-aware cross-frame correspondence. Section 3.3 introduces Streaming Grounding Reinforcement (SGR) with verifiable Grounding, Identity, and Consistency rewards. Section 3.4 presents the progressive three-stage training recipe. Section 3.5 summarizes data curation for single-frame and streaming supervision. Additional implementation details are deferred to the appendix.

3.1 Problem Formulation

Streaming visual grounding. Let q denote a free-form language query, either a category name (detection) or a referring expression (refer). Let $I_{1:T} = (I_1, \ldots, I_T)$ be a causally arriving RGB stream, where $I_t \in \mathrm{R}^{H \times W \times 3}$ is the observation at timestep t. At each timestep t, the model must

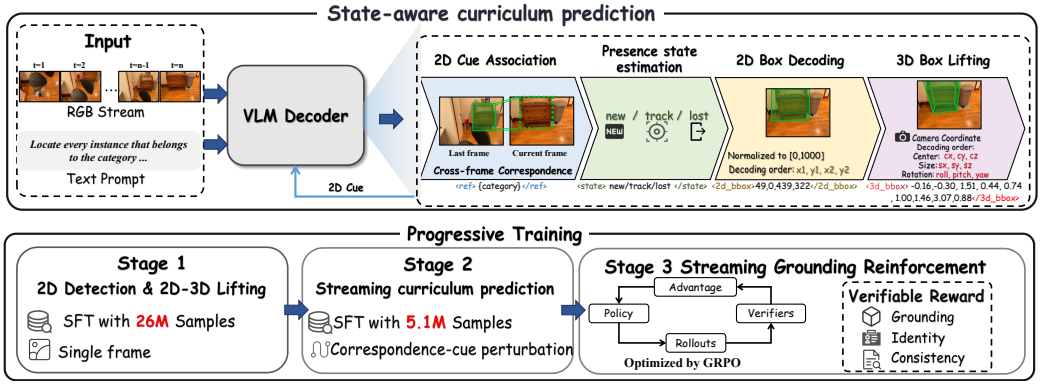

Figure 2: **TempoGround overview.** Top: state-aware curriculum prediction guided by cross-frame correspondence, consuming adjacent dual-image frames with the language query and previous 2D cue, from presence-aware 2D localization to camera-frame 3D lifting. Bottom: progressive three-stage training from single-frame SFT to streaming curriculum prediction and Streaming Grounding Reinforcement (SGR).

localize all objects in view that match q and output their 2D boxes and camera-frame 3D boxes. Formally, the prediction at timestep t is a (possibly empty) set

$$\mathcal{Y}_t = \left\{ y_t^{(k)} \right\}_{k=1}^{N_t}, \qquad y_t^{(k)} = \left(\mathbf{b}_t^{2\mathrm{D},(k)}, \mathbf{b}_t^{3\mathrm{D},(k)} \right), \tag{1}$$

where N_t is the number of predicted instances, $\mathbf{b}_t^{2\mathrm{D},(k)} \in \mathrm{R}^4$ is an axis-aligned 2D box, and $\mathbf{b}_t^{3\mathrm{D},(k)} \in \mathrm{R}^9$ is a camera-frame 3D box. The streaming setting is causal: the prediction at timestep t may depend only on q, observations $I_{1:t}$, and previous predictions $\mathcal{Y}_{1:t-1}$.

Geometry conventions. All 2D boxes are represented in normalized image coordinates $\mathbf{b}^{2\mathrm{D}} = (x_1, y_1, x_2, y_2)$ with each coordinate quantized to integers in $[0, 1000]$. All 3D boxes are expressed in the current camera frame as

$$\mathbf{b}^{3\mathrm{D}} = (c_x, c_y, c_z, s_x, s_y, s_z, r, p, y), \tag{2}$$

where (c_x, c_y, c_z) is the box center in meters, (s_x, s_y, s_z) is the size in meters, and (r, p, y) denotes roll, pitch, and yaw in radians.

3.2 Curriculum Prediction with State-Aware Cross-Frame Correspondence

TempoGround is built on a standard autoregressive VLM decoder. At timestep t, under the guidance of q, the decoder emits a structured token sequence that realizes the following curriculum: *2D correspondence → presence estimation → 2D box decoding → camera-frame 3D lifting*.

Cross-frame correspondence. Let $\hat{\mathcal{Y}}_{t-1}$ denote the model prediction at the previous timestep (empty when $t = 1$). At each timestep $t > 1$, TempoGround takes both the previous image I_{t-1} and the current image I_t as visual inputs. We construct a lightweight correspondence cue

$$\mathbf{c}_{t-1} = \left\{ \left(\ell_{t-1}^{(k)}, \mathbf{b}_{t-1}^{2\mathrm{D},(k)} \right) \right\}_k, \tag{3}$$

which retains only labels and 2D boxes from $\hat{\mathcal{Y}}_{t-1}$ and discards 3D boxes. Camera-frame 3D boxes are coupled to the viewpoint at the previous timestep; carrying them forward would inject stale geometric bias into the current frame. The 2D cue instead provides a view-stable association signal, after which TempoGround re-estimates 3D in the current camera frame. Conditioned on the adjacent image pair (I_{t-1}, I_t), language query q, and correspondence cue $\mathbf{c}_{t-1}$, TempoGround associates instances across frames by matching $\mathbf{c}_{t-1}$ against corresponding regions in I_t.

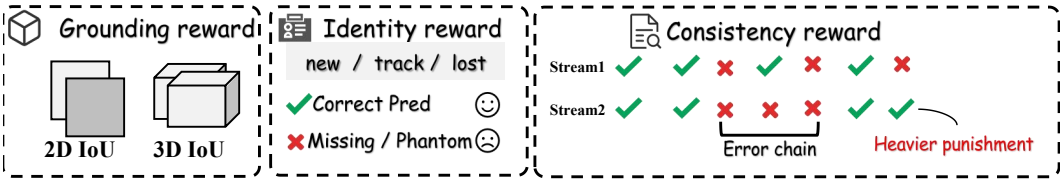

Figure 3: Verifiable rewards in SGR.

Presence state estimation. To strengthen cross-frame instance association, TempoGround predicts a discrete presence state for each associated instance. We use the presence vocabulary

$$\mathcal{S} = \{\texttt{new}, \texttt{track}, \texttt{lost}\}, \tag{4}$$

where `new`, `track`, and `lost` respectively indicate that an object newly enters the field of view, persists in the view, or exits the view. A predicted state $s_t \in \mathcal{S}$ determines whether the corresponding instance contributes a box to $\mathcal{Y}_t$. If a previously lost query-matching object re-enters the view, TempoGround treats it as `new` again: presence is defined with respect to contiguous visibility under language guidance, rather than as a persistent identity memory across exits. Supervising these states is critical for suppressing identity drift among currently visible instances and for preventing continued box predictions after objects leave the view.

From 2D boxes to camera-frame 3D. After correspondence and presence are resolved, TempoGround decodes a 2D box for each visible instance and then lifts it to a camera-frame 3D box of the form $\mathbf{b}^{\mathrm{3D}} = (c_x, c_y, c_z, s_x, s_y, s_z, r, p, y)$. Predicting 2D first and then lifting to 3D lets the 2D box serve as an intermediate spatial anchor that stabilizes monocular camera-frame 3D estimation. Multi-instance responses are serialized in a canonical order. The resulting autoregressive target interleaves presence, 2D, and 3D tokens in the same order used at inference.

3.3 Streaming Grounding Reinforcement

Token-level supervised fine-tuning trains the decoder to imitate discrete coordinate tokens. This objective cannot directly model continuous box geometry, nor does it provide trajectory-level supervision when each step depends on previous model outputs. We therefore introduce Streaming Grounding Reinforcement (SGR), an RLVR stage that optimizes TempoGround with verifiable rewards on streamed rollouts.

Objective. Given a streaming prompt $x = (q, I_{1:T})$, the policy π_θ samples a trajectory $\tau = \mathcal{Y}_{1:T}$. A deterministic verifier returns a scalar reward $R(\tau)$. SGR maximizes the expected reward

$$J(\theta) = \mathrm{E}_{x\sim\mathcal{D},\,\tau\sim\pi_\theta(\cdot|x)}\left[R(\tau)\right]. \tag{5}$$

We optimize this objective with Group Relative Policy Optimization (GRPO) Shao et al. (2024). For each prompt, GRPO samples a group $\{\tau_i\}_{i=1}^{G}$ from the old policy and computes the group-normalized advantage

$$A_i = \frac{R(\tau_i) - \mathrm{mean}(\{R(\tau_j)\}_{j=1}^{G})}{\mathrm{std}(\{R(\tau_j)\}_{j=1}^{G}) + \epsilon}, \tag{6}$$

where $\epsilon > 0$ is a small constant for numerical stability. During SGR rollouts, the correspondence cue $\mathbf{c}_{t-1}$ is taken from the model's own previous prediction, so optimization reflects closed-loop streaming behavior.

Verifiable rewards. SGR assigns each closed-loop trajectory a scalar reward

$$R(\tau) = R_{\mathrm{g}}(\tau) + R_{\mathrm{id}}(\tau) + R_{\mathrm{con}}(\tau) - f_{\mathrm{fmt}}(\tau), \tag{7}$$

where f_{fmt} penalizes unparsable outputs. As illustrated in Figure 3, The three main terms are:

- **Grounding** R_{g}: encourages accurate and persistent 2D/3D boxes for query-matching objects throughout the stream;

Table 1: Training sources for TempoGround. Stage 1 uses 15.6M 2D samples and 10.5M camera-frame 2D+3D samples. Stage 2 uses 5.1M samples, from which Stage 3 draws a subset.

Stage	Modality	Sources	Samples
1	2D	**Objects365** Shao et al. (2019), **V3Det** Wang et al. (2023), **COCO** Lin et al. (2014), **LVIS** Gupta et al. (2019), **EgoObjects** Zhu et al. (2023), **BDD100K** Yu et al. (2020), **nuImages** Caesar et al. (2020), **PACO** Ramanathan et al. (2023), **RefCOCO/+/g** Kazemzadeh et al. (2014); Yu et al. (2016), **RefL4** Chen et al. (2025), **Flickr30k Entities** Plummer et al. (2015)	15.6M
	2D+3D	**ScanNet++** Yeshwanth et al. (2023), **ScanNet** Dai et al. (2017), **ADT** Pan et al. (2023), **ASE** Avetisyan et al. (2024), **ARKitScenes** Baruch et al. (2021), **Hypersim** Roberts et al. (2021), **Objectron** Ahmadyan et al. (2021), **SUN-RGBD** Song et al. (2015), **KITTI** Geiger et al. (2012), **nuScenes**	10.5M
2–3	Stream	**ARKitScenes**, **ASE**, **ADT**, **ScanNet++**, **ScanNet**	5.1M

- **Identity** R_{id}: encourages correct enter and exit behavior, so objects are detected when they appear and are not predicted after they leave;
- **Consistency** R_{con}: discourages long streaks of failures that propagate through self-generated correspondence cues.

Computation details are provided in the appendix.

3.4 Progressive Three-Stage Training

We train TempoGround with a progressive three-stage curriculum that moves from single-frame geometric competence to streaming supervision and finally to sequence-level reinforcement.

Stage 1: Single-frame 2D detection and 2D-to-3D lifting. Stage 1 applies supervised fine-tuning on about 26M single-frame samples covering large-scale 2D detection/referring and camera-frame 3D detection/referring. The model learns open-vocabulary localization and the 2D-then-3D decoding order that underlies later curriculum prediction.

Stage 2: Streaming curriculum prediction. Stage 2 continues supervised fine-tuning on about 5.1M samples. Each sample provides an adjacent image pair (I_{t-1}, I_t), a language query, and a correspondence cue derived from the previous frame, and supervises presence states together with 2D/3D boxes. This stage installs state-aware cross-frame correspondence so that presence estimation and consistent localization become learnable under temporal cues.

Stage 3: SGR. Stage 3 refines the Stage 2 policy with SGR on streamed trajectories under the objective and rewards introduced above. This stage aligns training with geometric and consistency objectives that token-level imitation cannot capture.

3.5 Data Curation

We curate a camera-centric corpus that supports both single-frame Stage 1 supervision and streaming Stages 2–3. Heterogeneous public sources are converted into a unified JSONL schema and packaged as VLM conversations whose response order matches the curriculum above. Detailed filtering thresholds, packaging templates, and per-source statistics are provided in the appendix.

Sources and scale. Table 1 summarizes training sources. Stage 1 mixes large-scale 2D corpora with camera-frame 2D+3D corpora, totaling about 26M samples. Stages 2–3 are built from multi-view RGB-D sequences with stable object identities. Stage 2 uses 5.1M streaming samples, from which Stage 3 draws a subset for SGR.

Unified geometry interface. Every sample is stored in a shared JSONL format with image pointers, language queries, instance lists, and camera intrinsics. Box normalization and geometry con-

Table 2: Non-clean correspondence-cue modes in Stage 2 (80% of samples; the remaining 20% use clean GT cues). Each mode corrupts only the input prior $\mathbf{c}_{t-1}$; supervision targets remain unchanged.

Mode	Ratio	Role in training
Stage 1 rollout	25%	Uses frozen Stage 1 predictions as cues to expose realistic closed-loop errors.
Jitter	20%	Shifts prior 2D boxes to simulate fast ego-motion and cross-frame misalignment.
Dropout	12%	Removes prior instances to teach recovery from previous-step misses.
Phantom	12%	Inserts false boxes or wrong labels to discourage trusting spurious cues.
Stale	7%	Replaces the $t-1$ cue with an older cue to handle outdated association.
Empty	4%	Clears the cue to force re-initialization after a stream break.

ventions follow Section 3.1. This unification removes dataset-specific heads and keeps Stage 1 3D supervision consistent with streaming camera-frame prediction.

Quality control. Aggregating heterogeneous public sources inevitably introduces inconsistent category granularity and unreliable geometry. We canonicalize ambiguous or overly fine-grained labels, remove geometrically degenerate classes, and discard low-quality boxes through frustum, visibility, and consistency checks. Calibrated negatives are further included so that absent queries teach rejection rather than hallucination. On streaming sources, presence-state labels follow per-frame visibility, and clips are sampled with randomized length and stride to diversify temporal contexts. Concrete filtering thresholds and cleaning pipelines are provided in the appendix.

Correspondence-cue perturbation. A natural concern for cue-conditioned streaming is error amplification: once the previous prediction is wrong, under fast ego-motion the previous 2D box may misalign severely with the current image, and a model that over-relies on $\mathbf{c}_{t-1}$ can drift across subsequent frames. To prevent such brittle dependence, Stage 2 constructs $\mathbf{c}_{t-1}$ from a mixture of clean ground-truth cues, Stage 1 rollouts, and controlled synthetic perturbations, while keeping supervised targets unchanged. Table 2 lists the non-clean modes, which jointly cover 80% of Stage 2 samples; the remaining 20% use clean previous-frame ground-truth 2D boxes. Rollout cues provide realistic closed-loop errors; synthetic modes further cover ego-motion misalignment, missed and spurious detections, stale cues, and stream restarts. Implementation details of each operator are deferred to the appendix.

View-grounded referring annotation. Public datasets such as ScanRefer provide scene-level referring expressions that are not suited to map-free streaming, so we discard them as primary supervision and regenerate view-grounded queries on ego imagery. Each expression avoids absolute directional phrases tied to the global scene layout and instead relies on target attributes and relations to co-visible anchors in the current view. Along a stream, the target and its anchors jointly determine the first resolvable timestep: the object is labeled `new` only when the query becomes uniquely decidable from the observed view, after which presence states follow the target with cross-frame correspondence. Annotation protocols and uniqueness checks are detailed in the appendix.

4 Experiments

4.1 Implementation Details

Training details. We initialize from Qwen3.5 and fully fine-tune all parameters with AdamW in ms-swift Zhao et al. (2025). We use bfloat16, FlashAttention, and DeepSpeed ZeRO-3, with weight decay 0.05 and gradient clipping at 1.0. All training runs on 128 Alibaba Cloud PPUs and takes about 109 hours.

Stage 1 trains for 1 epoch with learning rate 4×10^{-5}, vision learning rate 2×10^{-5}, sequence length 8192, packing length 8192, and global batch size 1024, taking about 69 hours.

Table 3: 3D grounding results. Best results are in **bold**; second best are underlined.

Method	ARKitScenes		ScanNet		ScanNet++		Mean	
	F1@0.25	AP_{3D}	F1@0.25	AP_{3D}	F1@0.25	AP_{3D}	F1@0.25	AP_{3D}
Qwen3-VL-4B	14.3	8.4	19.2	9.3	14.5	8.1	16.0	8.6
Qwen3-VL-8B	19.4	11.2	21.2	11.2	14.4	12.4	18.3	11.6
Qwen3.5-4B	10.3	7.3	27.3	11.0	8.8	8.4	15.5	8.9
Qwen3.5-9B	14.8	9.9	22.2	11.6	11.9	9.2	16.3	10.2
VG-LLM	27.9	11.5	25.3	18.3	22.7	13.1	25.3	14.3
DetAny3D	21.0	14.4	51.3	30.6	31.3	18.5	34.5	21.2
OVMono3D	46.3	24.7	52.9	29.8	46.2	29.4	48.5	28.0
TempoGround-4B (ours)	47.6	25.5	56.1	38.6	40.7	27.5	48.1	30.5
TempoGround-9B (ours)	**49.5**	**30.8**	**63.0**	**45.3**	**51.7**	**30.4**	**54.7**	**35.5**

Table 4: 2D grounding results. Best results are in **bold**; second best are underlined.

Method	ASE		ARKitScenes		ScanNet++		Mean	
	F1@0.5	F1@0.95	F1@0.5	F1@0.95	F1@0.5	F1@0.95	F1@0.5	F1@0.95
Qwen3-VL-4B	36.8	5.1	36.0	11.9	54.1	19.2	42.3	12.1
Qwen3-VL-8B	38.0	2.4	34.4	11.1	54.2	17.7	42.2	10.4
Qwen3.5-4B	16.4	1.5	31.0	6.3	52.9	13.5	33.4	7.1
Qwen3.5-9B	12.1	1.7	33.7	7.1	56.4	12.7	34.1	7.2
Cosmos-Reason2-8B	12.4	2.1	33.7	6.3	43.3	7.6	29.8	5.3
GroundingDINO-T	43.4	20.1	45.7	13.8	74.5	**40.5**	54.5	24.8
GLaMM	13.5	2.1	15.9	4.0	35.5	10.2	21.6	5.4
RexOmni-3B	43.3	19.4	42.9	17.6	65.8	38.3	50.7	25.1
LocateAnything	35.1	**21.7**	**52.1**	18.7	72.3	37.3	53.2	25.9
TempoGround-4B (ours)	38.2	19.4	49.8	21.0	62.3	30.5	50.1	23.6
TempoGround-9B (ours)	**45.5**	18.6	51.9	**21.2**	**79.4**	39.3	**58.9**	**26.4**

Stage 2 uses the same schedule for 1 epoch with sequence length 16384, packing length 16384, and global batch size 1024, taking about 21 hours.

Stage 3 trains with GRPO for 2 epochs on about 4.1K streaming trajectories, with 32 prompts per step, learning rate 1×10^{-6}, vision learning rate 5×10^{-7}, KL coefficient β=0.01, G=4 rollouts at temperature 1.0, and up to 15 streaming turns, taking about 19 hours.

Evaluation setup. To ensure a fair comparison, we evaluate all methods under the same causal streaming protocol. Single-frame methods take each frame independently; video-based methods take the clip observed so far at timestep t and report only the current-frame prediction. Each dataset contributes 24 test sequences with equal detection and referring samples. We report $F1_{2D}$@0.5/@0.95 for 2D and $F1_{3D}$@0.25/AP_{3D} for 3D. These metrics are sufficient for our setting because streaming failures ultimately surface as localization errors: incorrect enter/exit or identity decisions cause misses and spurious boxes, while cross-frame inconsistency yields unstable matches over time. F1 averages per-frame precision and recall along the stream and therefore reflects whether presence and correspondence remain correct at typical timesteps. AP_{3D} aggregates predictions over the full trajectory and is more sensitive to persistent false positives and fragmented tracks under closed-loop updates. Together, they evaluate both step-wise grounding quality and stream-level stability without requiring separate state-only scores as primary metrics.

4.2 Main Results

Table 3 and Table 4 report 3D and 2D results under causal streaming inputs. TempoGround consistently surpasses both general-purpose VLMs and specialist grounders, indicating that state-aware curriculum prediction and SGR are effective for online localization rather than only for offline single-frame recognition. The 2D gains under the same protocol further show that the benefit is

Table 5: Ablation of curriculum prediction.

Configuration	AP_{3D}
w/o correspondence prior	37.7
2D+3D prior	41.8
w/o presence-state prediction	40.5
Ours	**42.3**

Table 6: Ablation of SGR rewards.

Configuration	AP_{3D}
w/o SGR	42.3
w/o Grounding reward R_{g}	42.7
w/o Identity reward R_{id}	44.1
w/o Consistency reward R_{con}	43.9
Ours	**45.3**

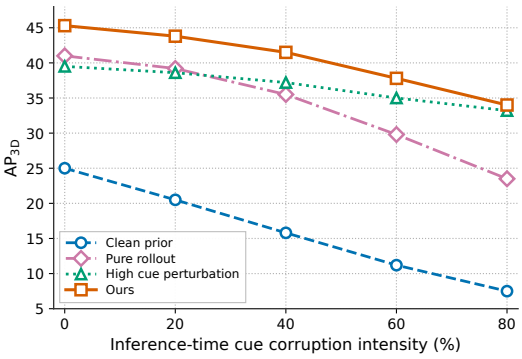

Figure 4: Closed-loop robustness to inference-time cue corruption on ScanNet (AP_{3D}).

not confined to camera-frame 3D lifting, but also improves image-plane grounding when predictions must be refreshed causally.

4.3 Ablation Studies

All ablations keep the same TempoGround-9B backbone, data budget, and evaluation protocol on ScanNet under causal streaming with AP_{3D}. Curriculum-prediction ablations use the Stage 2 supervised checkpoint for computational efficiency, while SGR ablations remove SGR or individual rewards from the full model.

Table 5 shows that both cross-frame correspondence and presence supervision improve AP_{3D}, to different degrees. Carrying camera-frame 3D boxes in the prior underperforms the 2D-only cue, consistent with Section 3.2: previous-view 3D geometry can bias localization under the current camera frame. Table 6 shows that Stage 3 SGR brings a clear overall gain, and removing any of Grounding, Identity, or Consistency lowers accuracy, indicating that all three rewards contribute positively.

4.4 In-depth Analysis

Under closed-loop streaming, errors in $\mathbf{c}_{t-1}$ can propagate across timesteps. We therefore corrupt the incoming cue at inference with spatial jitter, box dropout, and label mixing, sweeping the corruption intensity from 0% to 80%. Figure 4 compares four Stage 2 cue schedules on ScanNet AP_{3D}.

Training with clean ground-truth cues alone yields a low closed-loop starting point and a steep decay, revealing an oracle-prior gap. Pure Stage 1 rollout improves the clean operating point and remains relatively stable under mild corruption, but degrades once the noise exceeds the Stage 1 error distribution. High cue perturbation flattens the curve yet suppresses peak accuracy, indicating that excessive prior noise weakens useful association signals. *Ours*, which mixes rollout with the synthetic modes in Table 2, attains the best accuracy across the full intensity range: rollout supplies realistic closed-loop residuals, while synthetic perturbations cover rarer failure modes that rollout alone does not.

5 Conclusion

We presented TempoGround, a VLM-native framework for visual grounding on causally arriving image streams. Cross-frame correspondence and explicit presence states drive a curriculum from 2D association to camera-frame 3D grounding, and Streaming Grounding Reinforcement (SGR) optimizes that curriculum with verifiable Grounding, Identity, and Consistency rewards. TempoGround thus achieves accurate and consistent camera-frame grounding for streaming perception.

AI USE STATEMENT

We used Qwen3.7-Plus to propose candidate referring expressions during data curation. Accepted expressions were kept only after the programmatic and uniqueness checks in the appendix. Generative AI tools were also used to polish the manuscript, search and summarize related work, and assist with LaTeX formatting. The authors reviewed all AI-assisted text and annotations and take responsibility for the final content of this work.

REFERENCES

Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7822–7831, 2021.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision*, pp. 247–263. Springer, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 513–524, 2025.

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

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. *arXiv preprint arXiv:2510.12798*, 2025.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xiongkun Linghu, Jiangyong Huang, Baoxiong Jia, and Siyuan Huang. 3d-rft: Reinforcement fine-tuning for video-based 3d scene understanding. *arXiv preprint arXiv:2603.04976*, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yunze Man, Shihao Wang, Guowen Zhang, Johan Bjorck, Liang-Yan Gui, Jim Fan, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Locateanything3d: Vision-language 3d detection with chain-of-sight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31089–31102, 2026.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20133–20143, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pp. 51575–51598, 2024.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7141–7151, 2023.

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

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19844–19854, 2023.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, et al. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. *arXiv preprint arXiv:2605.27365*, 2026.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19757–19767, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026.

Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20110–20120, 2023.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.