G^2-OPD: GROUNDED AND GRASPABLE ON-POLICY DISTILLATION FOR DYNAMIC SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Compact vision-language models are attractive for embodied applications, yet transferring dynamic spatial reasoning from stronger teachers remains challenging. Fixed-trace distillation suffers from a mismatch between teacher-provided reasoning prefixes and the student's own inference trajectories, while outcome-based reinforcement learning provides sparse corrective signals when successful rollouts are rare. We therefore study on-policy distillation (OPD), which supervises student-generated trajectories with dense teacher distributions. However, standard OPD treats reasoning positions uniformly, ignoring whether a prediction is grounded in spatiotemporal evidence and whether the corresponding correction is learnable by the student. To this end, we introduce G^2-OPD, a Grounded and Graspable OPD framework that adaptively allocates teacher supervision along the student's trajectory. Grounding uses counterfactual video probes to measure sensitivity to visual content, temporal order, and motion. Graspability captures corrections that are both needed and learnable by combining teacher–student disagreement with teacher support for the student's leading candidates. We then combine these two signals through a regularized supervision-allocation rule that redistributes full-vocabulary reverse-KL supervision while preventing excessive concentration. These signals reweight full-vocabulary reverse-KL supervision without task-specific process rewards or additional inference-time modules. Ultimately, G^2-OPD achieves 51.38% accuracy on VLM4D and 34.49% accuracy on DSR-Bench, improving over the baseline by 11.02 and 12.12 percentage points, respectively.

1 INTRODUCTION

Dynamic spatial reasoning requires a model to track how objects and their relations evolve over time while disentangling object motion from viewpoint changes. This couples spatial structure, temporal order, and local motion across frames, demanding reasoning grounded in changing visual evidence (Chen et al., 2024a; Yang et al., 2025; Zhou et al., 2025b;a). For embodied applications with limited memory and computation, the key challenge is how to transfer this capability to compact vision-language models (VLMs).

Recent work has strengthened dynamic spatial understanding through dedicated data and post-training. MLLM-4D combines spatiotemporal instruction tuning with reinforcement fine-tuning, and 4DThinker trains models to reason through dynamic latent imagery (Yin et al., 2026; Chen et al., 2026). Related spatial VLMs also show that large-scale spatial data and structure-aware visual features can improve geometric reasoning from images and videos (Chen et al., 2024a; Wu et al., 2025). Together, these advances show how to strengthen dynamic spatial understanding, but they leave open how to transfer that capability from a stronger teacher to a smaller student. Crucially, the student must learn from the intermediate states encountered along its own reasoning trajectory, rather than only from teacher-generated solutions.

Fixed-trace distillation provides detailed teacher supervision, but trains the student on teacher-generated prefixes that may differ from the states it encounters at inference. This train–test mismatch motivates on-policy distillation (Agarwal et al., 2024; Ko et al., 2024), which instead supervises the student's own rollouts. Outcome-based reinforcement learning is also on-policy, but its final

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.

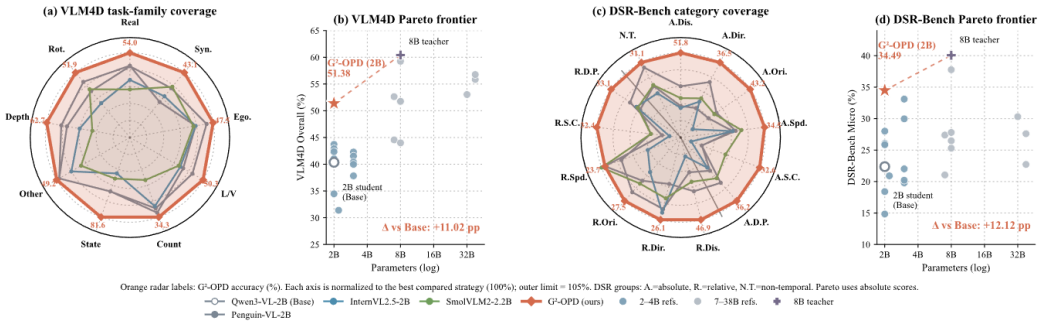


Figure 1: Performance overview on VLM4D and DSR-Bench. Radar plots show task-family and category coverage normalized to the best compared strategy (100%), while Pareto plots show the reported score against model size. The orange markers identify the 2B G^2-OPD model. Plotted values correspond to the benchmark tables.

correctness reward provides little information about which intermediate spatial judgment should be corrected. This is especially limiting when successful rollouts are rare and binary group-relative rewards exhibit little within-group variation. Dynamic spatial reasoning therefore calls for dense, trajectory-level feedback on the student's own states.

We therefore study OPD as a way to combine on-policy student trajectories with dense teacher supervision for dynamic spatial reasoning. At each generated prefix, the teacher provides a full next-token distribution, offering corrective feedback even when the final answer is wrong. However, not all trajectory states are equally informative for learning. We identify two properties that govern where teacher supervision is most useful: *grounding*, which reflects how strongly the teacher's prediction depends on spatial content, temporal order, and local motion; and *graspability*, which reflects whether a correction is both needed and learnable.

We introduce G^2-OPD, which adaptively redistributes dense teacher supervision along the student's own trajectory according to both *grounding* and *graspability*. Counterfactual video probes estimate grounding from sensitivity to spatial content, temporal order, and local motion, while local teacher–student disagreement and teacher support for student candidates estimate graspability. We combine these signals to redistribute full-vocabulary reverse-KL supervision within each response without altering the teacher target or the student's inference architecture. Experimental results show that the 2B student reaches 51.38% on VLM4D and 34.49% on DSR-Bench, outperforming the baseline by 11.02 and 12.12 points, respectively. Controlled analyses further validate the proposed supervision-allocation mechanism. Figure 1 summarizes this comparison across task-family/category coverage and model-scale Pareto views.

Our contributions are:

- We introduce G^2-OPD, an on-policy distillation framework for transferring dynamic spatial reasoning to compact VLMs. Unlike uniform OPD, G^2-OPD redistributes full-vocabulary teacher supervision across the student's generated trajectory based jointly on how strongly each state is grounded in dynamic visual evidence and how ready the student is to learn the corresponding correction.
- We develop a Grounding signal based on counterfactual video probes that measures how strongly teacher predictions depend on spatial content, temporal order, and local motion, enabling supervision to focus on states supported by dynamic visual evidence.
- We develop a Graspability signal that combines teacher support with local teacher–student disagreement over the student's leading candidates, identifying trajectory states where corrective supervision is both needed and learnable.
- Experimental results show that our method reaches 51.38% on VLM4D and 34.49% on DSR-Bench, outperforming the baseline by 11.02 and 12.12 points, respectively. Controlled analyses further validate the proposed supervision-allocation mechanism.

2 RELATED WORK

Dynamic spatial understanding. Recent benchmarks and spatial VLMs study motion, perspective, geometry, and temporal relations in videos and images (Zhou et al., 2025b;a; Yang et al., 2025; Chen et al., 2024a; Wu et al., 2025). MLLM-4D and 4DThinker strengthen this capability through spatiotemporal instruction data, latent visual reasoning, and reinforcement learning (Yin et al., 2026; Chen et al., 2026). We instead study how to transfer such capability from a stronger teacher to a compact student, deciding which student states should receive supervision and how much influence the teacher should exert.

On-policy distillation. Knowledge distillation and sequence-level distillation transfer teacher predictions to smaller models and generated outputs (Hinton et al., 2015; Kim and Rush, 2016). DAgger and recent distillation methods motivate learning on learner-induced states, reverse-KL objectives, and self-generated sequences (Ross et al., 2011; Ko et al., 2024; Agarwal et al., 2024; Gu et al., 2026). Video-OPD extends dense on-policy feedback to temporal video grounding (Li et al., 2026). Our method applies this transfer setting to evolving spatial relations, reference frames, and motion judgments.

Visual supervision and teachable positions. Recent methods use visual perturbations, gradient decomposition, capacity-aware targets, or teacher–student disagreement to identify useful supervision (Liu et al., 2026; Yoon et al., 2026; Xue et al., 2026; Mirzadeh et al., 2020; Wang et al., 2026). We combine these ideas into a response-local allocation rule: counterfactual spatial, order, and motion probes estimate grounding, while candidate support and local disagreement estimate graspability. The resulting weights preserve the original full-vocabulary teacher distribution while concentrating supervision where the student has evidence for, but has not yet resolved, the spatial judgment.

3 TEACHING DYNAMIC SPATIAL REASONING ON POLICY

Consider a video question asking whether a moving object approaches a reference object while the camera also moves. A student may identify both objects yet misinterpret their relative motion. OPD allows the teacher to provide feedback after the student's own description, including at the point where that interpretation begins to go wrong.

A shared prediction context. A training batch contains B video–question pairs $x_i = (v_i, u_i)$, where v_i is the video and u_i is the question. The student generates a response $y_i = (y_{i,1}, \ldots, y_{i,T_i})$ of T_i tokens using its current rollout parameters $\bar{\theta}$. At position t, the prefix $y_{i,<t}$ contains the tokens it has already generated. The teacher and student both receive the same context

$$c_{i,t} = (x_i, y_{i,<t}), \qquad p_{i,t}(a) = p_\theta(a \mid c_{i,t}), \quad q_{i,t}(a) = q_\phi(a \mid c_{i,t}). \tag{1}$$

Here a is any possible next token in the shared vocabulary $\mathcal{V}$. The student distribution $p_{i,t}$ uses the parameters θ being updated, while the teacher parameters ϕ remain fixed. The generated response is also fixed during this update. Thus the teacher supplies feedback for the student's actual prefix, rather than replacing it with a teacher-written explanation.

Learning from the teacher distribution. The teacher assigns a probability to every next token, providing more information than a single target token or a final correctness score. We train the student to match this distribution with reverse KL:

$$\ell_{i,t} = D_{\mathrm{KL}}(p_{i,t} \| q_{i,t}) = \sum_{a \in \mathcal{V}} p_{i,t}(a) \log \frac{p_{i,t}(a)}{q_{i,t}(a)}. \tag{2}$$

The sum covers the complete vocabulary at each prediction position. Standard OPD gives every active response position the same coefficient. With $N = \sum_{i=1}^{B} T_i$ denoting the total number of response tokens, its batch loss is

$$\mathcal{L}_{\mathrm{OPD}} = \frac{1}{N} \sum_{i=1}^{B} \sum_{t=1}^{T_i} \ell_{i,t}. \tag{3}$$

Gradients update the student distribution in Equation (2). A fresh rollout then produces the contexts for the next update. The remaining question is how strongly to teach each visited position; Section 4 develops that allocation. The prefix-distribution and binary-reward analyses are given in Section A.5.

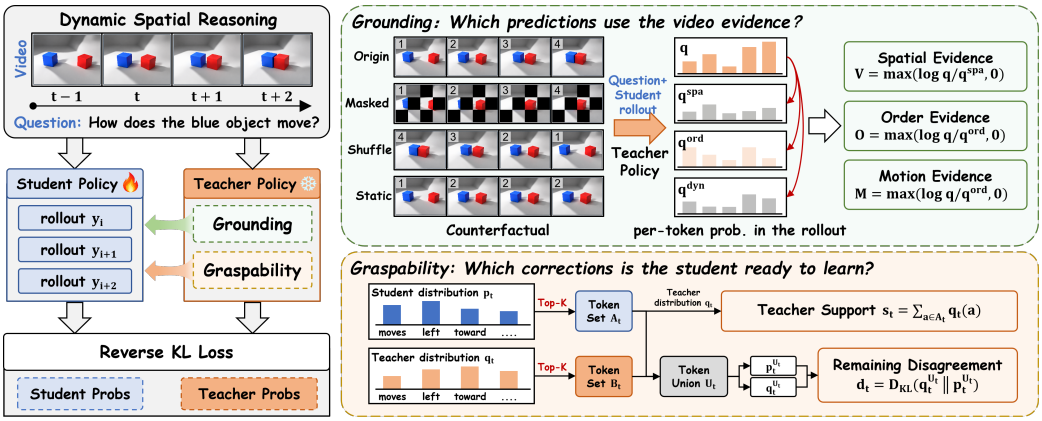


Figure 2: Overview of G^2-OPD. Student rollouts provide the shared prefixes. Grounding compares teacher probes under spatial, temporal-order, and local-motion counterfactuals, while graspability selects teacher-supported corrections among the student's top-K candidates. The two signals produce weights for the same full-vocabulary reverse-KL loss.

4 Grounded and Graspable On-Policy Distillation

Teacher feedback is dense, but its teaching value varies along a response. In the relative-motion example, identifying an object and deciding whether it approaches another involve different evidence and different levels of student readiness. G^2-OPD uses grounding to identify video-dependent predictions and graspability to identify corrections the student has a basis to learn. Their combination sets the weight on each OPD loss (as shown in Figure 2).

For the scoring definitions below, we discuss one response of length T and omit its batch index i. The subscript t always denotes a prediction position, and y_t is the token the student generated there. A score is computed separately at each position and then rescaled within that response.

4.1 Grounding: Which Predictions Use the Video Evidence?

To judge whether one object approaches another, a model must locate the objects, determine the direction of their changing relation, and follow motion between nearby observations. We probe these sources of evidence through paired intact and counterfactual video inputs, comparing teacher support at the same question and student-generated prefix (Liu et al., 2026).

The *spatial* counterfactual masks the same image regions across frames, weakening the visual cues used to identify objects and reference points while keeping the visibility pattern consistent through time. The *temporal-order* counterfactual reverses or locally reorders observations, testing judgments such as approach versus departure that depend on the direction of change. The *local-motion* counterfactual repeats a representative frame within short intervals. It retains the coarse sequence of scene states but suppresses the fine changes needed to estimate displacement, speed, and orientation. Construction settings are given in Section C.1.

Let q_t^{spa}, q_t^{ord}, and q_t^{dyn} denote the teacher distributions under the spatial, temporal-order, and local-motion controls. Comparing each with the intact-video distribution q_t gives three evidence scores at the student's generated token y_t:

$$V_t = [\log q_t(y_t) - \log q_t^{\mathrm{spa}}(y_t)]_+ , \tag{4}$$

$$O_t = \left[\log q_t(y_t) - \log q_t^{\mathrm{ord}}(y_t)\right]_+ , \tag{5}$$

$$M_t = \left[\log q_t(y_t) - \log q_t^{\mathrm{dyn}}(y_t)\right]_+ . \tag{6}$$

Here $[z]_+ = \max(z, 0)$. A positive score means that disrupting the corresponding evidence reduces teacher support for that token. Each contrast contributes separately, allowing supervision to respond to spatial identification, temporal direction, and local motion at different points in the reasoning trajectory.

For any one of these controls, indexed by $j \in \{\text{spa}, \text{ord}, \text{dyn}\}$, define $\delta_t^j(a) = \log q_t(a) - \log q_t^j(a)$. Its signed expectation under the student distribution satisfies

$$\mathbb{E}_{a\sim p_t}[\delta_t^j(a)] = D_{\mathrm{KL}}(p_t \| q_t^j) - D_{\mathrm{KL}}(p_t \| q_t). \tag{7}$$

Thus each contrast relates the evidence comparison to the same distribution-matching objective used in OPD. The practical scores retain its positive value at the sampled token; Section A.2 gives the derivation.

We rescale each score within the response and combine the three contributions:

$$g_t = \min\left\{ g_{\max},\, 1 + \alpha_{\text{spa}}\widehat{V}_t + \alpha_{\text{ord}}\widehat{O}_t + \alpha_{\text{dyn}}\widehat{M}_t \right\}. \tag{8}$$

The hats denote positive-score rescaling, and the nonnegative coefficients control the contribution of each evidence source. The unit baseline retains supervision at low-evidence positions, while $g_{\max}$ bounds the influence of any one position. The rescaling and configuration are specified in Section C.2 and Table 6.

4.2 Graspability: which corrections is the student ready to learn?

Even an evidence-dependent judgment can have different teaching value for different students. A student may already identify the relevant objects yet remain undecided about their relative motion. We call its readiness to learn the teacher's correction at this position *graspability*. Two quantities describe that readiness: how much the predictions still disagree, and how much the teacher supports next tokens the student already considers. We use the disagreement-and-support principle of TA-OPD (Wang et al., 2026) for soft, response-local weighting.

Let K denote the fixed candidate-set size, and let $\mathcal{T}_K(r)$ denote the set of the K highest-probability next tokens under distribution r. We write $A_t = \mathcal{T}_K(p_t)$ for the student candidates and $B_t = \mathcal{T}_K(q_t)$ for the teacher candidates. Teacher support is the total probability that the teacher assigns to the student set:

$$s_t = \sum_{a\in A_t} q_t(a). \tag{9}$$

To interpret s_t, consider the same reverse KL as in Equation (2), with predictions restricted to the student's candidates. The following result gives the best teacher match available within that region.

Proposition 1 (Candidate support and local matching)**.** *Fix a nonempty candidate set A and a teacher distribution q, and let $s = \sum_{a\in A} q(a) > 0$. If ν ranges over probability distributions that place all their mass on A, then*

$$\min_{\mathrm{supp}(\nu)\subseteq A} D_{\mathrm{KL}}(\nu\|q) = -\log s, \qquad \nu^*(a) = \frac{q(a)}{s} \quad (a \in A). \tag{10}$$

Here $\mathrm{supp}(\nu)$ contains tokens with positive probability under ν. At position t, the best candidate-restricted loss is $-\log s_t$: greater support means the current candidates can match the teacher more closely. This motivates the support factor in h_t. The auxiliary distribution ν evaluates the candidate region; the actual student loss remains full-vocabulary. Section A.3 gives the proof and its relation to that loss.

The candidate region's best fit and the student's remaining disagreement are different quantities. We measure disagreement on the union $U_t = A_t \cup B_t$, which includes leading tokens from both models. On this set, define the conditional distributions

$$p_t^{U_t}(a) = \frac{p_t(a)}{\sum_{b\in U_t} p_t(b)}, \qquad q_t^{U_t}(a) = \frac{q_t(a)}{\sum_{b\in U_t} q_t(b)} \quad (a \in U_t). \tag{11}$$

The superscript U_t means that each distribution is restricted to this set and normalized to sum to one. Local disagreement and the graspability multiplier are then

$$d_t = D_{\mathrm{KL}}(q_t^{U_t} \| p_t^{U_t}), \qquad h_t = 1 + \gamma \widehat{d}_t \widehat{s}_t. \tag{12}$$

The discrepancy d_t is zero when the two conditional distributions agree. The rescaled scores $\widehat{d}_t$ and $\widehat{s}_t$ lie in $[0, 1]$ and compare positions within the same response; γ controls the extra emphasis. Thus Proposition 1 motivates the support factor, while d_t identifies the correction still needed. Their product emphasizes teacher-supported continuations the student already considers but has not yet resolved. The baseline keeps every position supervised. Candidate sets are used to estimate this weight; the training loss still compares the full vocabulary.

4.3 COMBINING THE SIGNALS INTO ONE TEACHING WEIGHT

We now convert grounding and graspability into coefficients for the OPD loss. Standard OPD assigns a length-T response a total coefficient of T. We retain that total while directing more teaching toward high-scoring positions.

Let ρ_t be the fraction assigned to position t, giving $w_t = T\rho_t$, and let $u_t = 1/T$ be uniform allocation. The simplex $\Delta_T = \{\rho : \rho_t \geq 0, \sum_t \rho_t = 1\}$ contains all such fractions. The following objective favors both log scores and penalizes departure from uniform teaching.

Proposition 2 (Regularized supervision allocation)**.** *For fixed positive scores g_t, h_t and uniform $u_t = 1/T$, the allocation problem*

$$\rho^* = \arg\max_{\rho \in \Delta_T} \left\{ \sum_{t=1}^{T} \rho_t \big(\log g_t + \log h_t\big) - D_{\mathrm{KL}}(\rho \| u) \right\} \tag{13}$$

has the unique solution $\rho_t^* = g_t h_t / \sum_{t'} g_{t'} h_{t'}$.

This maximization chooses response-position fractions for fixed scores. We evaluate its closed-form normalized product, writing $r_t = g_t h_t$, then convert fractions to loss coefficients through $w_{i,t} = T_i \rho_{i,t}^*$:

$$w_{i,t} = \frac{g_{i,t} h_{i,t}}{T_i^{-1} \sum_{t'=1}^{T_i} g_{i,t'} h_{i,t'}}, \qquad \sum_{t=1}^{T_i} w_{i,t} = T_i. \tag{14}$$

Substituting these coefficients into the original OPD loss gives two equivalent expressions, with $N = \sum_i T_i$ and $\ell_{i,t}(\theta) = D_{\mathrm{KL}}(p_{i,t} \| q_{i,t})$:

$$\mathcal{L}_{\mathrm{G^2\text{-}OPD}}(\theta) = \frac{1}{N} \sum_{i=1}^{B} \sum_{t=1}^{T_i} \mathrm{sg}(w_{i,t})\, \ell_{i,t}(\theta) = \sum_{i=1}^{B} \frac{T_i}{N} \sum_{t=1}^{T_i} \mathrm{sg}(\rho_{i,t}^*)\, \ell_{i,t}(\theta). \tag{15}$$

Proposition 2 therefore supplies the coefficients of the student's matching objective. Stop-gradient (sg) holds those fractions fixed while minimizing over θ. The allocation KL compares response positions; the training KL compares vocabulary distributions. The factor T_i/N preserves the global token mean. Section A.4 derives the resulting student gradient and concentration bound. After the update, a new student rollout provides new prefixes and candidate sets. The weights therefore follow the student's changing spatial judgments, while the original teacher distribution continues to supervise every position. The counterfactual probes are used during training; the deployed student keeps its original inference architecture. Section B gives the complete training procedure.

5 EXPERIMENTS

We evaluate the dynamic spatial capability acquired by a student and examine how on-policy teaching, graspability, and grounding contribute to that transfer.

5.1 EVALUATION SETTING

We evaluate Qwen3-VL-2B on VLM4D and DSR-Bench (Zhou et al., 2025b;a). VLM4D covers real and synthetic videos across seven reported task families; DSR-Bench includes absolute and relative spatial judgments and a non-temporal category. We report sample-level accuracy (Overall or Micro) and category-average accuracy (Macro), in percent. The training comparison uses Base, Offline distillation, GRPO, Uniform OPD, OPD + Graspability, and G^2-OPD in the same student regime. Larger models provide cross-scale references; ranking marks compare the compact models or the nonteacher training methods. The evaluation protocols and scoring rules are summarized in Section D.

5.2 COMPARISON ACROSS BOTH BENCHMARKS

G^2-OPD achieves 51.38 Overall and 53.91 Macro on VLM4D, and 34.49 Micro and 35.05 Macro on DSR-Bench (Tables 1 and 2). These are the highest aggregate scores among the listed compact

Table 1: VLM4D comparison with source averages, seven task-family scores. Overall is the sample-level aggregate and Macro is the seven-family macro average. S marks spatial/4D-specialized models and the Teacher is shown at the top. Best and second-best values are marked within the 2–4B comparison group.

Model	Source avg. Real	Source avg. Syn.	Task families Ego.	L/V	Count	State	Other	Depth	Rot.	Avg. Overall	Avg. Macro
MLLM-4D-8B (Teacher)^{S}(Yin et al., 2026)	62.51	53.93	56.25	60.66	44.60	60.53	75.41	69.08	57.04	60.41	60.51
7–38B references											
Qwen2.5-VL-32B(Bai et al., 2025b)	50.98	59.33	48.75	54.40	42.72	73.68	59.02	57.80	41.48	53.03	53.98
InternVideo2.5-Chat-8B(Wang et al., 2025b)	24.51	24.72	21.25	23.65	17.84	44.74	13.11	36.71	11.85	24.56	24.16
LLaVA-NeXT-Video-7B(Zhang et al., 2024)	28.15	40.45	27.50	30.86	25.35	42.11	40.98	35.84	25.19	31.17	32.55
VG-LLM-QA-8B^{S}(Zheng et al., 2025)	47.70	32.58	50.00	43.37	28.17	57.89	54.10	50.00	45.93	44.00	47.07
VLM-3R-7B^{S}(Fan et al., 2026)	49.53	29.21	43.75	43.58	33.80	55.26	45.90	54.91	38.52	44.55	45.10
LLaVA-Video-7B(Zhang et al., 2025b)	50.33	39.78	46.25	46.45	39.91	60.53	45.90	57.23	42.96	47.74	48.46
InternVL2.5-8B(Chen et al., 2024b)	49.97	43.82	33.75	49.84	38.03	57.89	52.46	59.25	31.85	48.46	46.15
Qwen2.5-VL-7B(Bai et al., 2025b)	51.93	49.66	48.75	52.28	39.91	68.42	54.10	58.96	39.26	51.38	51.67
InternVL3.5-8B(Wang et al., 2025a)	54.05	44.49	48.75	49.73	43.19	57.89	57.38	60.40	54.07	51.71	53.06
Qwen3-VL-8B(Bai et al., 2025a)	50.69	55.06	45.00	51.54	39.91	60.53	59.02	63.29	40.74	51.76	51.43
2–4B models											
SmolVLM2-2.2B(Marafioti et al., 2025)	30.63	33.71	36.25	33.51	18.31	42.11	32.79	28.32	38.52	31.39	32.83
DeepSeek-VL2-Tiny(Wu et al., 2024)	31.51	39.10	42.50	31.81	23.47	47.37	52.46	39.60	25.93	33.37	37.59
InternVL2.5-2B(Chen et al., 2024b)	36.76	27.42	37.50	34.68	29.58	36.84	39.34	37.86	27.41	34.47	34.74
4DThinker-4DRL-3B^{S}(Chen et al., 2026)	39.53	32.58	40.00	36.06	27.23	42.11	44.26	45.66	41.48	37.83	39.54
Phi-3.5-Vision-Instruct(Abdin et al., 2024)	41.21	27.87	42.50	37.86	30.05	57.89	32.79	43.35	31.11	37.94	39.36
4DThinker-DiFT-3B^{S}(Chen et al., 2026)	42.89	31.01	40.00	37.96	36.62	44.74	39.34	44.51	46.67	39.98	41.41
TEMPURA-Qwen2.5-VL-3B-s2(Cheng et al., 2025)	42.23	34.83	45.00	41.25	30.05	68.42	50.82	40.75	34.81	40.42	44.44
VideoLLaMA3-2B(Zhang et al., 2025a)	43.25	34.16	43.75	38.92	38.03	47.37	52.46	45.66	40.00	41.02	43.74
CounterVid-Qwen2.5-VL-3B(Poppi et al., 2026)	44.71	31.69	47.50	39.02	34.74	65.79	54.10	46.82	40.00	41.52	46.85
Qwen2.5-VL-3B(Bai et al., 2025b)	45.37	32.81	**52.50**	40.51	32.86	52.63	**55.74**	47.69	40.74	42.29	46.10
Penguin-VL-2B(Zhang et al., 2026)	45.44	32.81	43.75	42.95	30.52	55.26	47.54	47.40	37.04	42.35	43.49
Qwen3-VL-2B(Bai et al., 2025a)	45.81	23.60	37.50	36.59	32.39	55.26	47.54	51.73	44.44	40.36	43.64
Ovis2.5-2B(Lu et al., 2025)	46.83	30.79	**52.50**	40.93	37.09	57.89	44.26	46.53	45.93	42.90	46.45
InternVL3.5-2B(Wang et al., 2025a)	46.32	35.73	38.75	41.46	**38.50**	57.89	50.82	51.45	43.70	43.72	46.08
TA-OPD (2B)(Wang et al., 2026)	51.86	38.20	46.25	46.87	29.11	76.32	52.46	60.12	**52.59**	48.51	51.96
G^2-OPD (2B)	**54.05**	**43.15**	47.50	**50.27**	34.27	**81.58**	49.18	**62.72**	51.85	**51.38**	**53.91**

Table 2: DSR-Bench comparison with absolute and relative task categories. A.=absolute, R.=relative, S.C.=speed comparison, D.P.=direction prediction, N.T.=Non-Temp. Micro is the sample-level aggregate; Macro is the category macro average.

Model	Task categories A.Dis.	A.Dir.	A.Ori.	A.Spd.	A.S.C.	A.D.P.	R.Dis.	R.Dir.	R.Ori.	R.Spd.	R.S.C.	R.D.P.	N.T.	Avg. Micro	Avg. Macro
MLLM-4D-8B (Teacher)^{S}(Yin et al., 2026)	51.81	52.38	61.73	48.81	46.74	43.97	45.95	46.59	52.50	17.53	30.22	30.12	30.21	40.07	42.96
7–38B references															
InternVL2.5-38B(Chen et al., 2024b)	22.89	28.57	20.99	33.33	7.61	26.72	14.41	29.55	26.25	27.84	5.04	19.28	32.77	22.72	22.71
InternVL3.5-38B(Wang et al., 2025a)	24.10	26.98	38.27	38.10	34.78	30.17	26.13	25.00	35.00	19.59	14.39	20.48	32.77	27.60	28.13
Qwen2.5-VL-32B(Bai et al., 2025b)	31.33	31.75	34.57	44.05	35.87	26.72	36.94	31.82	30.00	25.77	20.86	24.70	30.64	30.31	31.15
LLaVA-NeXT-Video-7B(Zhang et al., 2024)	15.66	15.87	8.64	13.10	7.61	12.93	15.32	22.73	13.75	15.46	9.35	9.64	22.98	14.56	14.08
InternVL2.5-8B(Chen et al., 2024b)	18.07	31.75	17.28	17.86	7.61	30.17	18.92	23.86	21.25	16.49	2.88	19.88	39.15	21.60	20.40
LLaVA-Video-7B(Zhang et al., 2025b)	25.30	22.22	19.75	38.10	18.48	23.28	24.32	18.18	18.75	23.71	20.14	19.28	30.21	23.62	23.21
Video-OPD-8B(Li et al., 2026)	34.94	15.87	39.51	30.95	18.48	31.90	20.72	25.00	30.00	17.53	12.95	14.46	35.74	25.30	25.23
Qwen3-VL-8B(Bai et al., 2025a)	27.71	26.98	41.98	34.52	19.57	30.17	21.62	22.73	21.25	20.62	10.07	25.30	37.02	26.48	26.12
InternVideo2.5-Chat-8B(Wang et al., 2025b)	26.51	23.81	17.28	34.52	30.43	29.31	18.92	17.05	23.75	19.59	21.58	30.12	40.85	27.32	25.67
VLM-3R-7B^{S}(Fan et al., 2026)	31.33	28.57	22.22	38.10	28.26	28.45	27.93	22.73	27.50	16.49	31.65	19.88	31.49	27.39	27.28
InternVL3.5-8B(Wang et al., 2025a)	36.14	31.75	30.86	27.38	30.43	33.62	20.72	22.73	28.75	20.62	17.27	24.70	35.32	27.80	27.71
2–4B models															
InternVL2.5-2B(Chen et al., 2024b)	18.07	17.46	7.41	21.43	4.35	17.24	10.81	23.86	15.00	10.31	4.32	21.08	18.30	14.84	14.59
Penguin-VL-2B(Zhang et al., 2026)	19.28	14.29	17.28	21.43	8.70	18.97	19.82	22.73	23.75	18.56	11.51	19.28	21.28	18.40	18.22
Qwen2.5-VL-3B(Bai et al., 2025b)	16.87	9.52	16.05	21.43	9.78	23.28	18.02	25.00	16.25	18.56	7.19	23.49	31.91	19.79	18.26
CounterVid-Qwen2.5-VL-3B(Poppi et al., 2026)	18.07	14.29	13.58	22.62	5.43	18.97	14.41	28.41	13.75	20.62	7.19	27.11	34.89	20.21	18.41
SmolVLM2-2.2B(Marafioti et al., 2025)	24.10	19.05	20.99	25.00	18.48	23.28	25.23	19.32	20.00	**24.74**	11.51	20.48	21.70	20.91	21.07
DeepSeek-VL2-Tiny(Wu et al., 2024)	25.30	19.05	18.52	28.57	13.04	20.69	18.92	23.86	20.00	17.53	18.71	16.87	29.79	21.39	20.83
Phi-3.5-Vision-Instruct(Abdin et al., 2024)	21.69	17.46	19.75	22.62	13.04	27.59	21.62	23.86	22.50	13.40	15.83	20.48	31.91	21.95	20.90
TEMPURA-Qwen2.5-VL-3B-s2(Cheng et al., 2025)	14.46	30.16	13.58	29.76	7.61	31.90	16.22	26.14	18.75	20.62	1.44	23.49	**37.45**	22.02	20.89
Qwen3-VL-2B(Bai et al., 2025a)	31.33	26.98	22.22	22.62	8.70	25.86	30.63	14.77	18.75	23.71	7.19	24.10	28.94	22.37	21.98
VideoLLaMA3-2B(Zhang et al., 2025a)	21.69	25.40	23.46	19.05	20.65	31.03	30.63	30.68	27.50	18.56	21.58	22.89	33.19	25.85	25.10
Ovis2.5-2B(Lu et al., 2025)	42.17	22.22	35.80	**45.24**	8.70	24.14	40.54	15.91	30.00	23.71	12.95	18.67	28.51	26.06	26.81
InternVL3.5-2B(Wang et al., 2025a)	25.30	25.40	19.75	38.10	23.91	31.03	25.23	30.68	26.25	23.71	20.14	30.12	34.89	28.01	27.27
4DThinker-DiFT-3B^{S}(Chen et al., 2026)	39.76	28.57	25.93	41.67	16.30	30.17	39.64	29.55	26.25	21.65	24.46	**36.75**	28.09	29.97	29.91
4DThinker-4DRL-3B^{S}(Chen et al., 2026)	49.40	**36.51**	33.33	41.67	30.43	34.48	**49.55**	**39.77**	**35.00**	23.71	25.18	28.92	24.26	33.10	34.79
TA-OPD (2B)(Wang et al., 2026)	27.71	30.16	14.81	30.95	19.57	28.45	23.42	14.77	23.75	12.37	26.62	24.70	24.68	23.48	23.23
G^2-OPD (2B)	**51.81**	**36.51**	**43.21**	34.52	**32.61**	**36.21**	46.85	26.14	27.50	23.71	**32.37**	33.13	31.06	**34.49**	**35.05**

Table 3: Comparison of training strategies on VLM4D and DSR-Bench.

Method	VLM4D				DSR-Bench	
	Real	Syn.	Overall	Macro	Micro	Macro
Teacher (reference)	62.51	53.93	60.41	60.51	40.07	42.96
Base (Qwen3-VL-2B)	45.81	23.60	40.36	43.64	22.37	21.98
Offline distillation	45.22	37.08	43.23	44.61	25.71	26.49
GRPO(Shao et al., 2024)	45.51	36.40	43.28	45.58	26.69	27.50
Uniform OPD	48.80	38.20	46.20	47.12	28.64	29.18
G^2-OPD	**54.05**	**43.15**	**51.38**	**53.91**	**34.49**	**35.05**

Table 4: Ablation of grounding and graspability weighting on VLM4D and DSR-Bench.

Method	VLM4D				DSR-Bench	
	Real	Syn.	Overall	Macro	Micro	Macro
Teacher (reference)	62.51	53.93	60.41	60.51	40.07	42.96
Base (Qwen3-VL-2B)	45.81	23.60	40.36	43.64	22.37	21.98
Uniform OPD	48.80	38.20	46.20	47.12	28.64	29.18
Grounded-OPD	52.88	37.75	49.17	50.84	31.92	32.21
Graspable-OPD	51.71	38.43	48.46	51.42	31.29	31.56
G^2-OPD	**54.05**	**43.15**	**51.38**	**53.91**	**34.49**	**35.05**

models. Relative to the base Qwen3-VL-2B, sample-level accuracy improves by 11.02 and 12.12 points. The next-highest compact sample-level scores are 48.51 on VLM4D and 33.10 on DSR-Bench, yielding margins of 2.87 and 1.39 points. The same student architecture therefore supports stronger spatial reasoning after targeted transfer.

The 2B model is also competitive with selected larger references. On VLM4D, its Overall is within 0.38 points of Qwen3-VL-8B, while its Macro is higher (53.91 versus 51.43); it exceeds LLaVA-Video-7B by 3.64 Overall points. On DSR-Bench, it exceeds Qwen2.5-VL-32B by 4.18 Micro points and InternVL3.5-38B by 6.89. These comparisons show that task-directed teacher supervision can bring a compact model close to, or beyond, considerably larger models on the reported dynamic spatial evaluations.

5.3 Gains on evolving spatial judgments

At the task level, the updated comparison tables show strong compact-model scores on VLM4D State (81.58), Ego. (50.00), and Depth (62.72), and on DSR-Bench A.Dir. (36.51), A.Ori. (43.21), and A.S.C. (32.61). The source averages on VLM4D are 54.05 on real videos and 43.15 on synthetic videos. These results place the transfer gains in the changing relations that motivate the method: the student must identify objects, order observations, and track local change. The combined teaching rule strengthens this task-specific capability while retaining the compact model's inference architecture.

5.4 From demonstrations to on-policy teaching

Uniform OPD improves sample-level accuracy over offline distillation on both benchmarks: 43.23 to 46.20 on VLM4D and 25.71 to 28.64 on DSR-Bench (Table 3). These gains support teaching on the learner's own prefixes as an effective transfer route. Allocating that feedback with G^2-OPD further raises the scores to 51.38 and 34.49, for gains of 8.15 and 8.78 points over offline distillation. The full method also exceeds GRPO by 8.10 and 7.80 points. Together, these comparisons show the value of coupling dense student-trajectory supervision with a rule that selects where the teacher should exert greater influence.

5.5 Graspability and grounding improve teaching allocation

Graspability adds 2.26 Overall points on VLM4D and 2.65 Micro points on DSR-Bench over Uniform OPD; Macro improves by 4.30 and 2.38 points (Table 4). The gains under both aggregation schemes support the use of student readiness when assigning dense feedback. Teacher-supported candidates and unresolved disagreement provide a more useful teaching criterion than equal coefficients at every position. Adding grounding raises the graspability variant from 48.46 to 51.38 on VLM4D and from 31.29 to 34.49 on DSR-Bench, with further Macro gains of 2.49 and 3.49 points. The aggregate gains support the complementary roles of the two signals: graspability identifies corrections the student has a basis to learn, while grounding directs that teaching toward predictions supported by the video.

5.6 Mechanism analysis of the allocation

We analyze a collection of fixed-completion replays to test whether the two allocation signals select the positions described in Sections 4.1 and 4.2. The mechanism overview in Figure 3 shows selective channel responses: spatial and order probes are strongest for directional judgments, while motion

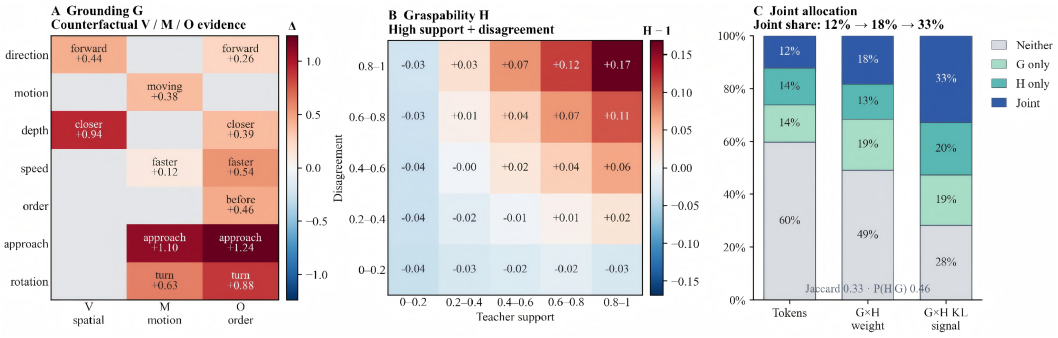


Figure 3: Overview of the supplied mechanism analysis. Panel A reports semantic-by-channel evidence, Panel B shows the support–disagreement geometry, and Panel C shows the partial overlap and concentration of the joint allocation. All three panels use the full confirmation pool.

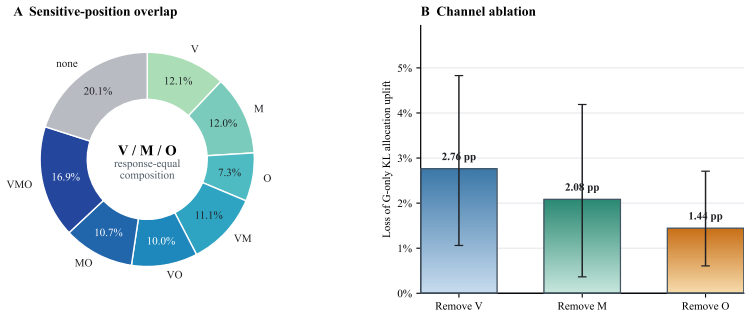


Figure 4: V/M/O sensitive-position overlap and channel ablation on a collection of response positions. Panel A partitions response positions by positive raw spatial (V), motion (M), and order (O) evidence. Panel B reports the decrease in G-only KL allocation uplift after removing one channel and recomputing the response-local weight.

probes emphasize movement-related predictions. The pattern is content-dependent rather than a uniform rescaling of all response positions.

Grounding and graspability overlap only partially, consistent with their distinct roles: grounding measures dependence on counterfactual evidence, whereas graspability measures teacher-supported disagreement. As summarized in Figure 4, spatial, motion, and order channels contain both complementary and shared sensitive positions; removing any one channel lowers the G-only allocation uplift. This supports using all three evidence channels in the grounding score.

At the token level, dynamic-space words receive higher grounding scores when the corresponding evidence is available, and the graspability multiplier is largest where teacher support and student disagreement are jointly high. The resulting weights concentrate KL supervision on positions that are both video-dependent and locally learnable, while preserving the response-level allocation budget.

6 Conclusion

We study how a stronger VLM can teach dynamic spatial reasoning to a compact student. OPD provides dense feedback at the student's own generated prefixes. G^2-OPD then allocates that feedback according to its dependence on spatiotemporal evidence and the student's readiness to learn the associated correction. Teacher support under these counterfactual probes identifies evidence-dependent predictions, while disagreement and candidate support estimate graspability. The analysis connects these signals to local distribution matching and characterizes the response-wise weights as a regularized allocation with bounded concentration. This formulation makes the teaching process explicit: the teacher supervises the student's evolving spatial understanding, with greater influence at grounded judgments that remain within its current predictive grasp.

REPRODUCIBILITY STATEMENT

Section A provides complete proofs of the analytical results. Sections B and C specify the training algorithm, score normalization, gradient flow, and distributed reduction. Section D describes the evaluation and diagnostic protocols. The source package contains the manuscript, result macros, an evidence ledger, and numerical checks of the mathematical identities.

AI USE STATEMENT

We used a large language model to assist with language editing and polishing of the manuscript, including improving grammar, clarity, and readability. All scientific content, experimental design, analysis, and conclusions were developed and verified by the authors, who take full responsibility for the paper.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024a. URL https://arxiv.org/abs/2401.12168.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xiang An, Bo Li, Xin Xie, ZiDong Wang, Mingze Sun, Shuang Chen, Hongyu Li, Xiaobin Hu, and Ruqi Huang. 4DThinker: Thinking with 4D imagery for dynamic spatial understanding. *arXiv preprint arXiv:2605.05997*, 2026. URL https://arxiv.org/abs/2605.05997.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Jen-Hao Cheng, Vivian Wang, Huayu Wang, Huapeng Zhou, Yi-Hao Peng, Hou-I Liu, Hsiang-Wei Huang, Kuang-Ming Chen, Cheng-Yen Yang, Wenhao Chai, Yi-Ling Chen, Vibhav Vineet, Qin Cai, and Jenq-Neng Hwang. Tempura: Temporal event masked prediction and understanding for reasoning in action. *arXiv preprint arXiv:2505.01583*, 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31054–31065, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2026. URL https://arxiv.org/abs/2306.08543v6. Version 6; first posted in 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of EMNLP*, 2016. URL https://arxiv.org/abs/1606.07947.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 24872–24895, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-OPD: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *arXiv preprint arXiv:2602.02994*, 2026. URL https://arxiv.org/abs/2602.02994.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026. URL https://arxiv.org/abs/2605.21924.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of AAAI*, 2020. URL https://arxiv.org/abs/1902.03393.

Tobia Poppi, Burak Uzkent, Amanmeet Garg, Lucas Porto, Garin Kessler, Yezhou Yang, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Florian Schiffers. CounterVid: Counterfactual video generation for mitigating action and temporal hallucinations in video-language models. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of AISTATS*, 2011. URL https://arxiv.org/abs/1011.0686.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025b.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026. URL https://arxiv.org/abs/2605.26844.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. URL https://arxiv.org/abs/2505.23747.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun,

Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill what the student can see: Fisher-projected on-policy distillation for vision-language models. *arXiv preprint arXiv:2608.01263*, 2026. URL https://arxiv.org/abs/2608.01263.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Xingyilang Yin, Chengzhengxu Li, Jiahao Chang, Chi-Man Pun, and Xiaodong Cun. MLLM-4D: Towards visual-based spatial-temporal intelligence. *arXiv preprint arXiv:2603.00515*, 2026. URL https://arxiv.org/abs/2603.00515.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026. URL https://arxiv.org/abs/2606.00564.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Boqiang Zhang, Lei Ke, Ruihan Yang, Qi Gao, Tianyuan Qu, Rossell Chen, Dong Yu, and Leoweiliang. Penguin-vl: Exploring the efficiency limits of vlm with llm-based vision encoders. *arXiv preprint arXiv:2603.06569*, 2026.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. LLaVA project blog.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4D: Dynamic spatial understanding for vision language models. *arXiv preprint arXiv:2512.20557*, 2025a. URL https://arxiv.org/abs/2512.20557.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: Towards spatiotemporal awareness in vision language models. In *Proceedings of ICCV*, 2025b. URL https://arxiv.org/abs/2508.02095.

A NOTATION AND PROOFS

We use natural logarithms and finite next-token vocabularies. Teacher and student softmax distributions are strictly positive. Results about a single prediction position condition on its video, question, and generated prefix. All scoring statistics and rollout prefixes are fixed during the gradient update under study.

The appendix follows the teaching decisions in Section 4: evidence contrasts (Section A.2), candidate support and its gradient connection (Section A.3), and joint allocation with bounded concentration (Section A.4). The context-distribution and outcome-feedback calculations for Section 3 follow in Section A.5.

A.1 NOTATION AT A GLANCE

Symbol	Meaning
i, t, B, T_i, N	Response index, position index, batch size, response length, and total response tokens $N = \sum_i T_i$.
$x_i = (v_i, u_i)$	A video and its question.
$y_{i,t}, y_{i,<t}$	Generated token at position t, and the student text preceding it.
$c_{i,t}$	Shared prediction context $(x_i, y_{i,<t})$; distinct from the multiplier h_t.
$a, \mathcal{V}$	A possible next token and the shared output vocabulary.
$p_{i,t}, q_{i,t}$	Student and intact-video teacher next-token distributions at the same context.
$q_t^j, \delta_t^j(a)$	Counterfactual teacher distribution and its log-support contrast; $j \in \{\mathrm{spa}, \mathrm{ord}, \mathrm{dyn}\}$.
V_t, O_t, M_t	Spatial, temporal-order, and local-motion support contrasts at the generated token.
$\widehat{V}_t, \widehat{O}_t, \widehat{M}_t$	Each evidence channel rescaled separately within the response.
A_t, B_t, U_t, K	Student's K leading next tokens, teacher's K leading next tokens, and their union.
$q_t^{U_t}, p_t^{U_t}$	Conditional distributions renormalized on U_t.
s_t, d_t	Teacher mass on A_t and local teacher-to-student disagreement on U_t.
$\widehat{s}_t, \widehat{d}_t$	Unit-range rescaling of each statistic within a response.
g_t, h_t, w_t	Grounding multiplier, graspability multiplier, and final mean-one teaching weight.
$\alpha_{\mathrm{spa}}, \alpha_{\mathrm{ord}}, \alpha_{\mathrm{dyn}}$	Nonnegative coefficients of the three evidence channels.
$\gamma, g_{\max}$	Graspability coefficient and grounding cap; configuration is in Table 6.
$\nu, \mathrm{supp}(\nu)$	A candidate-restricted probability distribution in Proposition 1, and the tokens to which it assigns positive mass.
$r_t, \rho_t, u_t, \Delta_T$	Raw priority $g_t h_t$, allocated fraction w_t/T, uniform fraction $1/T$, and all probability allocations over T positions.
$\ell_{i,t}, \mathrm{sg}$	Full-vocabulary reverse-KL loss and the stop-gradient operation.

Table 5: Symbols used in the teaching process and its analysis. Index i is omitted when discussing a single response.

A.2 GROUNDING: FIXED-PREFIX EVIDENCE CONTRASTS

For any intact-video teacher q, counterfactual teacher q^j, and student p,

$$D_{\mathrm{KL}}(p\|q^j) - D_{\mathrm{KL}}(p\|q) = \sum_a p(a)\left[\log\frac{p(a)}{q^j(a)} - \log\frac{p(a)}{q(a)}\right] \tag{16}$$

$$= \sum_a p(a)\log\frac{q(a)}{q^j(a)}. \tag{17}$$

Thus the untruncated log-support ratio has the expectation stated in Equation (7). A token drawn from p supplies a sample of this ratio; the method then takes its positive part for weighting. If rollout sampling uses a different temperature or truncation policy, the expectation uses that sampling distribution instead. The identity applies to that distribution as well.

Apply this identity separately to q^{spa}, q^{ord}, and q^{dyn}. The three implemented contrasts are

$$E_t^j = \left[\log q_t(y_t) - \log q_t^j(y_t)\right]_+, \qquad j \in \{\mathrm{spa}, \mathrm{ord}, \mathrm{dyn}\}, \tag{18}$$

with $(E_t^{\mathrm{spa}}, E_t^{\mathrm{ord}}, E_t^{\mathrm{dyn}}) = (V_t, O_t, M_t)$. Each channel compares the intact teacher with one counterfactual probe at the fixed prefix. For arbitrary nonnegative coefficients, the signed contrasts satisfy

$$\mathbb{E}_{a\sim p_t}\left[\sum_j \alpha_j \delta_t^j(a)\right] = \sum_j \alpha_j \left[D_{\mathrm{KL}}(p_t\|q_t^j) - D_{\mathrm{KL}}(p_t\|q_t)\right]. \tag{19}$$

This follows by linearity of expectation. The training weights use the positive, response-rescaled contrasts from these comparisons. Their separate rescaling lets spatial, order, and local-motion evidence each contribute to g_t at a given position.

A.3 Graspability: candidate support and local correction

Proof of Proposition 1. Let $s = q(A) > 0$ and define $q_A(a) = q(a)/s$ for $a \in A$. Every distribution ν supported on A satisfies

$$D_{\mathrm{KL}}(\nu\|q) = \sum_{a\in A} \nu(a) \log \frac{\nu(a)}{s q_A(a)} \tag{20}$$

$$= D_{\mathrm{KL}}(\nu\|q_A) - \log s. \tag{21}$$

The first term is nonnegative and vanishes uniquely at $\nu = q_A$. This proves both the minimum and its optimizer. □

At a scoring position, substitute $A = A_t$ and $q = q_t$ to obtain $-\log s_t$. This is the support-restricted version of the reverse-KL matching problem in Equation (2). It motivates the support factor used to form h_t; the actual student distribution remains defined on the complete vocabulary. The next decomposition connects the restricted comparison to that full student loss. The same interpretation admits an exact decomposition for the unrestricted student. Let $m = p(A)$, $s = q(A)$, and let p_A, q_A and $p_{\bar{A}}, q_{\bar{A}}$ denote conditional distributions within and outside A. Let $\mathrm{Bern}(m)$ denote the two-outcome distribution with probabilities m and $1-m$. For $0 < m, s < 1$,

$$D_{\mathrm{KL}}(p\|q) = D_{\mathrm{KL}}\big(\mathrm{Bern}(m)\|\,\mathrm{Bern}(s)\big) + m D_{\mathrm{KL}}(p_A\|q_A) + (1-m) D_{\mathrm{KL}}(p_{\bar{A}}\|q_{\bar{A}}). \tag{22}$$

To verify the result, substitute $p(a) = m p_A(a)$ and $q(a) = s q_A(a)$ on A, perform the analogous substitution on its complement, and collect terms. The first term measures disagreement about the probability mass assigned to the candidate region. The remaining terms measure disagreement within the two regions. Logging $m = p(A)$ alongside $s = q(A)$ therefore helps interpret how well the top-K set represents the student's predictions.

A.3.1 Reverse-KL gradient connection

Let z_a be the student logit for token a, and let z collect all vocabulary logits. For $p = \mathrm{softmax}(z)$,

$$\frac{\partial D_{\mathrm{KL}}(p\|q)}{\partial z_a} = p(a)\left(\log\frac{p(a)}{q(a)} - D_{\mathrm{KL}}(p\|q)\right). \tag{23}$$

Let $\ell = \sum_b p(b)\log[p(b)/q(b)]$ and $p = \mathrm{softmax}(z)$. The indicator $\mathbf{1}\{a = b\}$ equals one when $a = b$ and zero otherwise. The softmax derivative is $\partial p(b)/\partial z_a = p(b)(\mathbf{1}\{a = b\} - p(a))$. Applying the chain rule,

$$\frac{\partial \ell}{\partial z_a} = \sum_b p(b)(\mathbf{1}\{a = b\} - p(a))\left(\log\frac{p(b)}{q(b)} + 1\right) \tag{24}$$

$$= p(a)\left(\log\frac{p(a)}{q(a)} + 1\right) - p(a)(\ell + 1) \tag{25}$$

$$= p(a)\left(\log\frac{p(a)}{q(a)} - \ell\right). \tag{26}$$

For fixed positive q on a finite vocabulary, ℓ is bounded above by $\log(1/\min_b q(b))$. Both $p(a)\log p(a)$ and $p(a)(-\log q(a) - \ell)$ tend to zero as $p(a) \to 0$. For $p = \mathrm{softmax}(z/\tau)$ the derivative acquires a factor $1/\tau$. The training loss here does not introduce an additional τ^2 multiplier.

A.4 JOINT ALLOCATION AND COEFFICIENT CONCENTRATION

Proof of Proposition 2. Let $r_t > 0$, $u_t = 1/T$, and $Z = \sum_t u_t r_t$. Define $\rho_t^* = u_t r_t / Z$. For any $\rho \in \Delta_T$, where Δ_T consists of nonnegative vectors whose entries sum to one,

$$\sum_t \rho_t \log r_t - D_{\mathrm{KL}}(\rho \| u) = \sum_t \rho_t \log \frac{u_t r_t}{\rho_t} \tag{27}$$

$$= \log Z - D_{\mathrm{KL}}(\rho \| \rho^*). \tag{28}$$

The unique maximum is attained at $\rho = \rho^*$. Since u is uniform, $\rho_t^* = r_t / \sum_\tau r_\tau = w_t / T$. Moreover, $\log r_t = \log g_t + \log h_t$, giving the additive evidence and graspability utilities in the variational expression. □

To connect this allocation to the student update, keep the rollout, scores, and teacher distributions fixed. Since $w_{i,t} = T_i \rho_{i,t}^*$ and $N = \sum_i T_i$, substitution into the original weighted loss yields both forms of Equation (15). Differentiating only through the student distribution gives

$$\nabla_\theta \mathcal{L}_{\text{G}^2\text{-OPD}} = \sum_{i=1}^{B} \frac{T_i}{N} \sum_{t=1}^{T_i} \rho_{i,t}^* \, \nabla_\theta D_{\mathrm{KL}}(p_{i,t} \| q_{i,t}). \tag{29}$$

Thus the fractions solving the allocation problem are exactly the coefficients of the student's per-position learning gradients, after accounting for response length. The allocation KL compares positions; the KL in this gradient compares vocabulary distributions. The closed-form allocation is recomputed on the next rollout and remains detached within each update.

A.4.1 CONCENTRATION OF THE SUPERVISION WEIGHTS

Corollary 1 (Bounded concentration)**.** *For the configured $g_{\max} = 2$ and $\gamma = 0.25$, the effective number of weighted positions satisfies*

$$N_{\text{eff}} := \frac{(\sum_t w_t)^2}{\sum_t w_t^2} \geq \frac{40}{49} T \approx 0.816T. \tag{30}$$

Proof of Corollary 1. The evidence scores are nonnegative, so the implemented grounding weight lies in $[1, 2]$. Each unit score lies in $[0, 1]$, giving $h_t \in [1, 1 + \gamma] = [1, 1.25]$. Therefore $r_t = g_t h_t \in [a, b] = [1, 2.5]$.

For any $r \in [a, b]$, $(r - a)(r - b) \leq 0$, hence $r^2 \leq (a + b)r - ab$. Writing $\bar{r} = T^{-1} \sum_t r_t$, we obtain

$$\frac{T^{-1} \sum_t r_t^2}{\bar{r}^2} \leq \frac{(a + b)\bar{r} - ab}{\bar{r}^2} \leq \frac{(a + b)^2}{4ab}. \tag{31}$$

The last bound follows by maximizing the middle expression over $\bar{r} \in [a, b]$; its maximum occurs at $\bar{r} = 2ab/(a + b)$. Scaling all weights by the same constant leaves N_{eff} unchanged, so

$$\frac{N_{\text{eff}}}{T} = \frac{\bar{r}^2}{T^{-1} \sum_t r_t^2} \geq \frac{4ab}{(a + b)^2} = \frac{40}{49}. \tag{32}$$

□

For a general nonnegative graspability coefficient γ, the same calculation gives $b = 2(1 + \gamma)$ and $N_{\text{eff}}/T \geq 4b/(1 + b)^2$. The quantity N_{eff} describes coefficient concentration across positions; the gradients at different positions may still have different magnitudes and directions.

A.5 ON-POLICY CONTEXTS AND OUTCOME FEEDBACK

Write μ_S for the distribution of contexts $c = (x, y_{<t})$ reached by the student and μ_O for the contexts in a fixed offline trace set. Let $f(c)$ be a per-context diagnostic bounded between zero and $L_{\max}$. Their expected diagnostics differ by at most

$$|\mathbb{E}_{\mu_S} f - \mathbb{E}_{\mu_O} f| \leq L_{\max} D_{\mathrm{TV}}(\mu_S, \mu_O), \qquad D_{\mathrm{TV}}(P, Q) = \sup_C |P(C) - Q(C)|. \tag{33}$$

Here D_{TV} is total variation distance, and the supremum ranges over measurable sets of contexts C.

Proof. Write $f(c) = \int_0^{L_{\max}} \mathbf{1}\{f(c) > u\}\, du$. Integrating the difference between the two probabilities of the event $\{f > u\}$ and applying the definition of total variation yields the bound. In a discrete context space, the same result follows by summing the positive and negative parts of $\mu_S - \mu_O$. □

The premise specifies the class of diagnostics to which the bound applies. For example, a clipped per-context discrepancy is bounded. For next-token KL itself, a uniform positive lower bound on teacher probabilities supplies a finite bound. OPD samples the student prefix distribution directly at each rollout iteration.

For one question, let π be the probability that a rollout succeeds and let K be the number of independently sampled rollouts. A group has identical binary rewards exactly when all rollouts succeed or all fail. These events have probabilities π^K and $(1-\pi)^K$, respectively. Hence the probability of obtaining both reward outcomes is

$$P_{\text{mixed}} = 1 - \pi^K - (1-\pi)^K. \tag{34}$$

At $\pi = 0.01$ and $K = 8$, this is 0.0772553, or approximately 7.73%. This calculation concerns the differentiating signal from the binary task reward. Additional reward terms or training regularizers have their own contributions.

B Training algorithm

Algorithm 1 Grounded and Graspable On-Policy Distillation

Require: Student p_θ, frozen teacher q_ϕ, training inputs, $K = 16$
Require: Evidence coefficients $\alpha_{\text{spa}}, \alpha_{\text{ord}}, \alpha_{\text{dyn}} \geq 0$; $\gamma = 0.25$
for each training minibatch $\{x_i = (v_i, u_i)\}$ **do**
  Generate responses y_i with the current student
  Encode the same responses for both models; align response tokens and prefixes
  Compute student distributions $p_{i,t}$ with gradients
  Construct spatial-mask, temporal-order, and local-motion teacher inputs
  Compute frozen distributions $q_{i,t}, q_{i,t}^{\text{spa}}, q_{i,t}^{\text{ord}}, q_{i,t}^{\text{dyn}}$
  Initialize weighted sum $S_{\text{loc}} \leftarrow 0$, token count $N_{\text{loc}} \leftarrow 0$
  for each nonempty response i **do**
    Select the T_i active shifted response positions
    Compute $\ell_{i,t} = D_{\text{KL}}(p_{i,t} \| q_{i,t})$ over the full vocabulary
    Gather teacher log probabilities at the generated token $y_{i,t}$
    Compute $V_{i,t}, O_{i,t}, M_{i,t}$ using Equations (4) to (6)
    Rescale each evidence channel within the response; add its contribution to $g_{i,t}$
    Form detached student and teacher top-K sets and their union
    Compute teacher support $s_{i,t}$ and local disagreement $d_{i,t}$
    Quantile-scale $s_{i,t}$ and $d_{i,t}$; compute $h_{i,t}$
    $w_{i,t} \leftarrow \text{sg}\big(g_{i,t} h_{i,t} / \text{mean}_t(g_{i,t} h_{i,t})\big)$
    $S_{\text{loc}} \leftarrow S_{\text{loc}} + \sum_t w_{i,t} \ell_{i,t}$; $N_{\text{loc}} \leftarrow N_{\text{loc}} + T_i$
  end for
  All-reduce N_{loc} to obtain the global active response-token count N
  Backpropagate $R S_{\text{loc}} / N$, where R is the DDP world size
  Apply the optimizer step after DDP gradient averaging
end for

The teacher distribution is retained as the target at every prediction position. The two scoring branches determine detached scalar weights. Computation of the KL can be chunked over response positions, and each chunk retains all vocabulary entries. Generated text supplies the next iteration's on-policy prefixes.

Intermediate normalizations. The implementation normalizes the grounding and graspability branches separately before multiplying them. Let $\bar{g} = T^{-1} \sum_t g_t$ and $\bar{h} = T^{-1} \sum_t h_t$. Their final

combination is

$$\frac{(g_t/\bar{g})(h_t/\bar{h})}{T^{-1}\sum_\tau (g_\tau/\bar{g})(h_\tau/\bar{h})} = \frac{g_t h_t}{T^{-1}\sum_\tau g_\tau h_\tau}. \tag{35}$$

This cancellation gives the compact expression used in the main text and in Algorithm 1.

C Implementation and reproducibility

C.1 Distribution alignment and counterfactual video probes

Both models score the same student-generated response at aligned, shifted next-token positions. The response mask determines which positions enter the loss and its denominator. The model pair uses a shared output-token vocabulary.

Counterfactual video construction. A video contains L logical frames on a fixed spatial patch grid. We construct three counterfactual teacher inputs in addition to the intact video.

Spatial evidence. We mask a fixed 50% subset of spatial patches across all frames.

Temporal order. We apply either full-sequence reversal or within-window shuffling using four-frame windows.

Local motion. We replace the frames within each window by its center frame, preserving changes between windows while suppressing local motion.

All counterfactual inputs preserve the question, generated text prefix, video length, and visual-token count. The exact sampling choices are fixed by the training configuration.

C.2 Scoring and numerical operations

For a nonnegative response score vector $z = (z_1, \dots, z_T)$, let $P = \{t : z_t > 0\}$ be its positive positions and $|P|$ their number. If P is empty, $\mathrm{PosNorm}(z) = 0$. Otherwise, $Q_{0.95}$ denotes the empirical 95th percentile and supplies the clipping threshold κ:

$$\kappa = Q_{0.95}(\{z_t : t \in P\}), \quad \widetilde{z}_t = \min(z_t, \kappa), \quad \mathrm{PosNorm}(z)_t = \frac{\widetilde{z}_t}{|P|^{-1}\sum_{t' \in P} \widetilde{z}_{t'}}. \tag{36}$$

Here $\widetilde{z}_t$ is the clipped score. Apply this operation separately to V, O, and M within every response:

$$\widehat{V}_t = \mathrm{PosNorm}(V)_t, \qquad \widehat{O}_t = \mathrm{PosNorm}(O)_t, \qquad \widehat{M}_t = \mathrm{PosNorm}(M)_t. \tag{37}$$

Each branch is rescaled using its own positive values before the three contributions are added in Equation (8). For the graspability statistics, $Q_{0.05}(z)$ and $Q_{0.95}(z)$ are the empirical 5th and 95th percentiles across response positions. The operation $\mathrm{clip}(z, 0, 1)$ limits values to $[0, 1]$. Define

$$\mathrm{Unit}(z)_t = \mathrm{clip}\left(\frac{z_t - Q_{0.05}(z)}{Q_{0.95}(z) - Q_{0.05}(z)}, 0, 1\right), \qquad \widehat{d}_t = \mathrm{Unit}(d)_t, \quad \widehat{s}_t = \mathrm{Unit}(s)_t. \tag{38}$$

A quantile range at most 10^{-8} gives a zero unit score. We use $K = 16$ and $\gamma = 0.25$ for graspability. Percentiles are computed over the active positions of each response. The positive-score normalization balances the three evidence scales; the unit map makes disagreement and teacher support comparable before multiplication. The unit baselines in g_t and h_t retain all-token supervision. These are the rescaling operations denoted by hats in Section 4.

The grounding multiplier is computed with the nonnegative coefficients α_{spa}, α_{ord}, and α_{dyn}, and capped at $g_{\max} = 2$. Consequently, $g_t \in [1, 2]$, and the coefficient-concentration result in Corollary 1 applies to the three-channel rule. The evidence coefficients are recorded with the training configuration in Table 6. Teacher support uses the original teacher probabilities summed over the student's top-K set. Local disagreement uses the deduplicated union of the student's and teacher's top-K sets, with each distribution renormalized on that union. This branch uses detached logits and float32 log-softmax computation. The full-vocabulary loss uses the temperature-scaled distributions $p = \mathrm{softmax}(z^S/\tau)$ and $q = \mathrm{softmax}(z^T/\tau)$. The scoring branch uses its stated untempered logits.

The experimental temperature and rollout sampling settings are fields in the run record; at $\tau = 1$ the distribution notation in the main text applies directly to both branches.

Component	Supplied method configuration
Base distillation	Full-vocabulary reverse KL; GKD coefficient 1
Teacher target	Original teacher distribution; target strength 1.0
Grounding	Separate spatial, order, and local-motion evidence
Evidence coefficients	$\alpha_{\mathrm{spa}}, \alpha_{\mathrm{ord}}, \alpha_{\mathrm{dyn}}$: 0.25
Grounding cap	$g_{\max} = 2$
Graspability	Soft weighting, $K = 16$, $\gamma = 0.25$
Quantile normalization	Positive-score $Q_{0.95}$; unit-score $Q_{0.05}, Q_{0.95}$
Degenerate unit range	Threshold 10^{-8}
Visual mask	50% spatial patches; one deterministic mask shared across frames
Order branch	Sample-level $1/2$ reversal, $1/2$ local shuffle
Temporal windows	Four frames for local shuffle and local static
Weight normalization	Mean one within each response
Batch reduction	Global active response-token mean
Gradient path	Student distribution in the KL term

Table 6: Method parameters specified by the supplied implementation description.

C.3 Global token reduction

Let R be the number of distributed data-parallel (DDP) ranks. Rank r has weighted loss sum S_r and active response-token count N_r. The target loss is

$$\mathcal{L} = \frac{\sum_{r=1}^{R} S_r}{N}, \qquad N = \sum_{r=1}^{R} N_r. \tag{39}$$

DDP averages the gradients from the R ranks. Each rank therefore backpropagates RS_r/N, using the all-reduced token count N. Their average is $\sum_r \nabla S_r/N$, the gradient of the global token-mean loss. Gradient-accumulation microbatches use the same token accounting when their lengths differ. A response of length T_i has coefficient sum T_i. Its relative contribution therefore follows its active-token count. The method requires one student forward pass for differentiable distributions and four frozen teacher conditions in addition to rollout generation. Teacher conditions are evaluated sequentially to limit peak memory.

C.4 Executable verification in the source package

The source package includes numerical checks of the fixed-prefix identities, the support projection, the logit derivative, the variational allocation, and the concentration bound. These checks exercise synthetic probability distributions and verify the mathematics and result arithmetic. Benchmark evaluation is specified separately in Section D; its run records populate the experimental tables.

D Evaluation protocols

The evaluations below connect each empirical claim to an observable comparison. We train a cold-started Qwen3-VL-2B student under the supervision of MLLM-4D-8B using the MLLM4D-R1-30k dataset. All comparisons follow the official VLM4D and DSR-Bench evaluation protocols, with the same designated splits and answer-scoring implementation.

D.1 Transfer through student-generated prefixes

Tables 3 and 4 compare the Qwen3-VL-2B base student, offline distillation, GRPO, Uniform OPD, Grounded-OPD, Graspable-OPD, and G2-OPD on VLM4D and DSR-Bench. Uniform OPD improves over offline distillation from 43.23 to 46.20 on VLM4D and from 25.71 to 28.64 on DSR-Bench. G2-OPD reaches 51.38 and 34.49, exceeding GRPO by 8.10 and 7.80 points and offline distillation by 8.15 and 8.78 points on the two benchmarks, respectively. These comparisons show the gains from supervising student-generated prefixes and allocating teacher feedback across the trajectory.

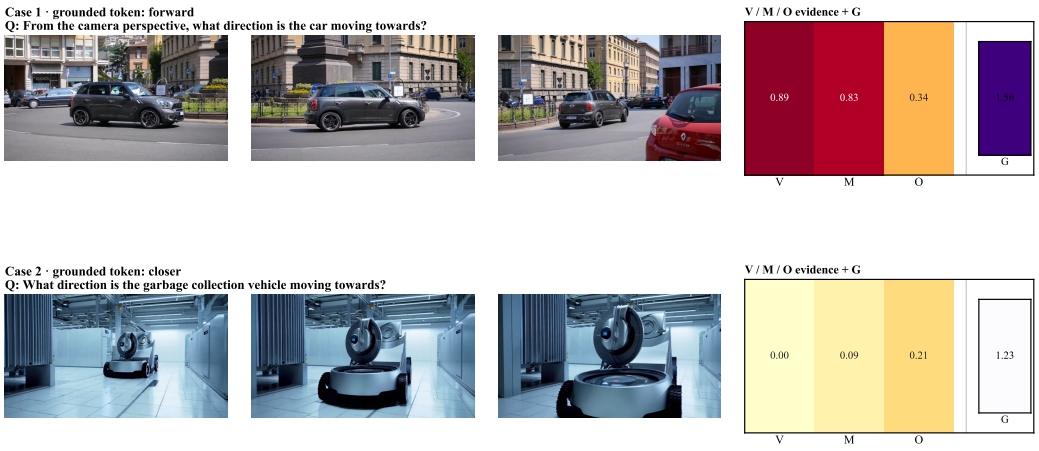


Figure 5: Grounding case studies for two motion judgments. Three original-video frames establish the motion context; V, M, and O share an absolute evidence scale, while G is shown separately. The forward case receives strong spatial and motion evidence, whereas the closer case is supported mainly by motion and order.

D.2 Two-factor and allocation controls

Table 4 reports all four cells of the grounding–graspability comparison: Uniform OPD (S_0), Grounded-OPD (S_G), Graspable-OPD (S_H), and G2-OPD (S_{GH}). Using the primary aggregate scores, the interaction term

$$\Delta_{\text{int}} = S_{GH} - S_G - S_H + S_0 \tag{40}$$

is -0.05 percentage points on VLM4D Overall and -0.08 points on DSR-Bench Micro. The corresponding Macro interactions are -1.23 and $+0.46$ points. The ablation results therefore quantify the separate and joint contributions of grounding and graspability under the reported evaluation setting.

D.3 Task-specific analysis

The main tables report the seven VLM4D families—Ego., L/V, Count, State, Other, Depth, and Rot.—together with Real/Syn. source averages. The evaluator-to-family mapping is recorded with the run metadata. DSR-Bench reports twelve absolute/relative temporal categories and N.T.; we retain its Micro and Macro aggregates. The temporal-only diagnostic averages the twelve displayed temporal scores and excludes N.T., and is reported separately from the official aggregates. Qualitative examples show sampled video frames, the question, a student prefix, teacher and student leading continuations, $V_t, O_t, M_t, d_t, s_t, w_t$, and the completed answer. Examples cover reference-frame selection, relative motion, and changes in object relations, with each example labeled by its actual prediction. This links the weighting mechanism to a concrete spatial judgment.

E Mechanism diagnostics

The mechanism section reports response-level diagnostics and token-level case studies. We analyze frozen replays with response-level aggregation and bootstrap intervals. Raw V, M, and O are positive log-support drops, while G and H are response-local normalized weights. The diagnostics proceed from channel evidence to grounding, readiness-based redistribution, and joint concentration.

The channel diagnostics in Figure 4 show the intended specialization across evidence channels. Direction and trajectory words respond to both spatial and order controls; motion-state words respond to the local-motion control; distance/depth words are strongest under the spatial control; and temporal-order and rotation words are strongest under the order control. Overlap states make the same point at the token level: V only, M only, and O only account for 12.06%, 11.97%, and 7.29% of tokens,

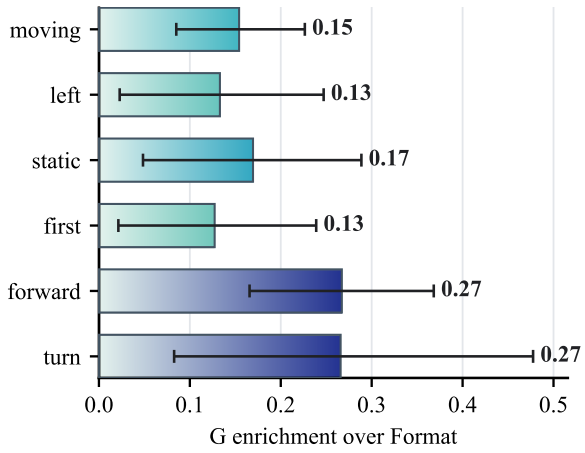


Figure 6: Grounding enrichment for complete dynamic-space words relative to matched Format tokens. Bars are response-equal estimates and black intervals are 95% bootstrap CIs.

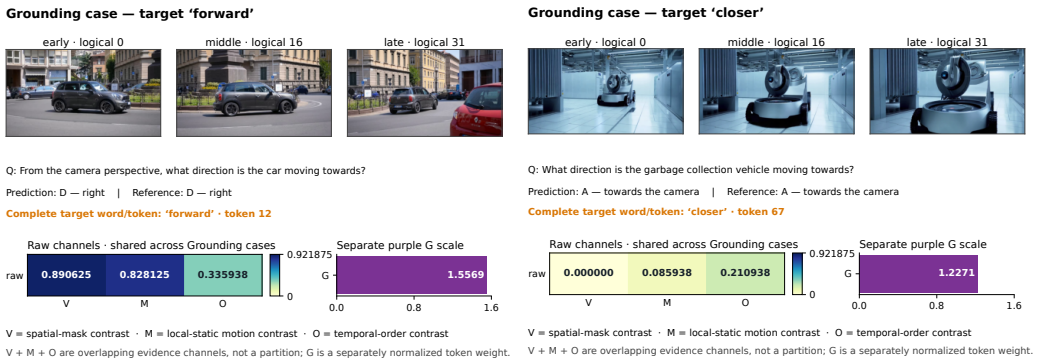


Figure 7: Grounding cases for the target words *forward* and *closer*. The complete words receive different combinations of spatial, motion, and order evidence. In the *forward* case, the target word appears in the reasoning trace while the answer is the direction "right".

whereas $V+M+O$ accounts for 16.92% and no positive channel for 20.07%. These shares are response-equal descriptive proportions.

Removing one channel lowers the full-system G-only allocation uplift by 2.762 percentage points for V, 2.084 for M, and 1.445 for O, with the corresponding 95% intervals shown in Figure 4. We report these as conditional allocation diagnostics, with each channel recomputed separately.

Figure 6 locates the same signal in complete words rather than tokenizer fragments. The displayed dynamic-space words all have positive enrichment with intervals above zero. These word-level results provide a readable view of the grounding signal.

In the *forward* case, all three raw channels are positive ($V = 0.890625$, $M = 0.828125$, $O = 0.335938$) and $G = 1.556922$. In the *closer* case, the positive evidence is mainly M and O (0, 0.085938, and 0.210938), with $G = 1.227129$. Together, the cases show how the multiplier responds to different evidence combinations.

The left panel of Figure 8 shows that H increases only when teacher support and student disagreement rise together. The high-support, high-disagreement cell has mean $H - 1 = +0.1686$, while high support with low disagreement remains negative. In the right panel, Q1–Q4 release KL share and Q5 absorbs $+2.063$ percentage points (95% CI $[1.803, 2.324]$); the response-level changes sum to approximately zero. This is the conservation property induced by mean-one normalization.

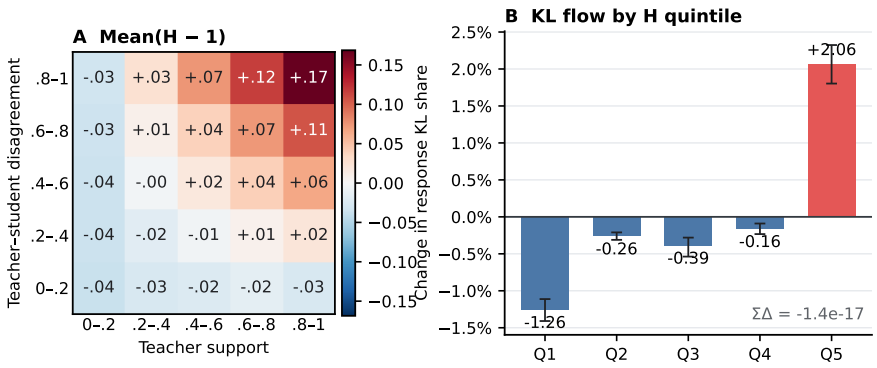


Figure 8: Graspability diagnostics. The left panel reports mean $H - 1$ across teacher-support and student-disagreement bins; the right panel reports the response-level KL-share change after sorting positions by H.

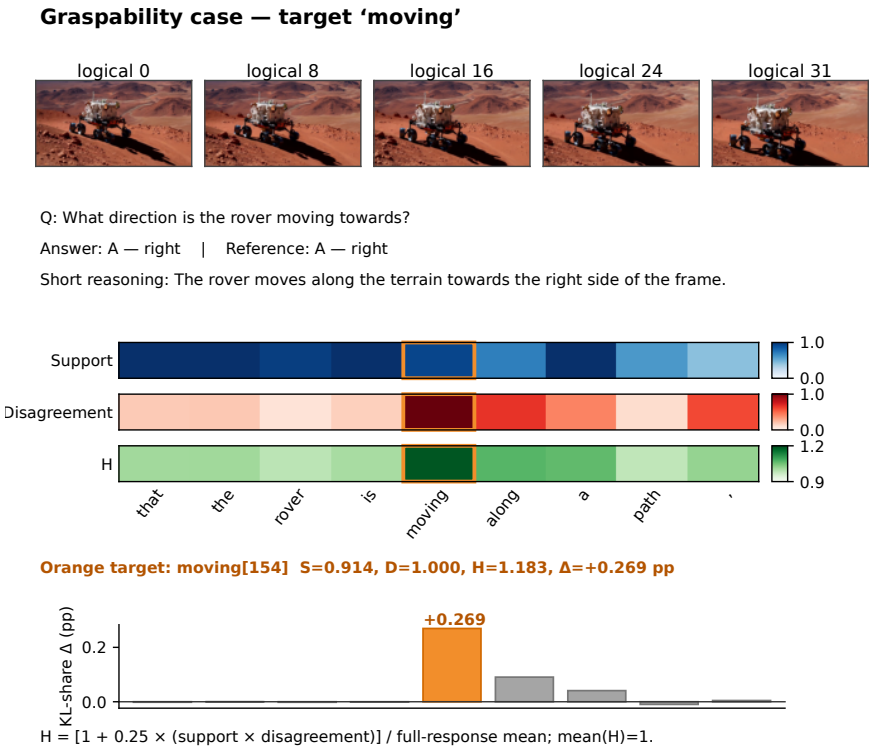


Figure 9: Qualitative graspability case for the target word *moving*. Teacher support and student–teacher disagreement jointly increase the response-local H weight and redistribute reverse-KL share within the response.

The *moving* example illustrates the selection rule summarized by the aggregate heatmap: support alone is insufficient when disagreement is low, while teacher support and unresolved disagreement together increase the local teaching weight.

The training trace shows stable optimization and persistent positive channel signals.

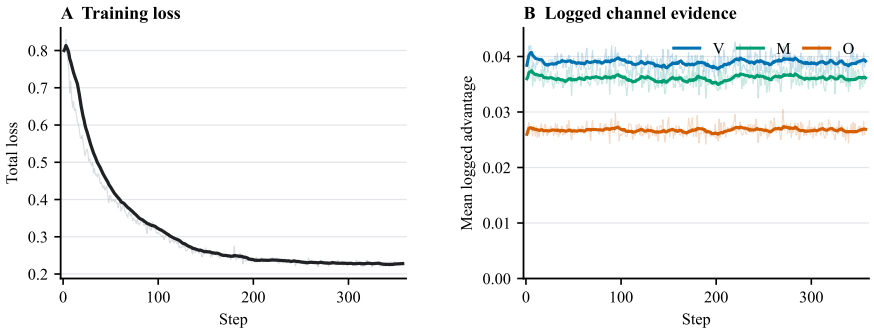


Figure 10: Training trajectory of total loss and logged V, M, and O means.