OPD-TV: RETHINKING TEACHER VIEWS FOR PRIVILEGED SUPERVISION TRANSFER IN VIDEO ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Privileged on-policy distillation extends on-policy distillation by allowing a teacher model to construct supervision from temporal evidence unavailable to the student. In Video OPD, this privilege is instantiated through additional temporal evidence available to the teacher while the student continues to operate under its original visual input. We find a supervision transfer gap: continuously enriching the teacher's visual view improves the teacher's grounding accuracy yet produces non-monotonic student gains. This supervision transfer gap arises because an enriched teacher can exploit visual cues unavailable to the student, introducing target shifts that the student cannot effectively absorb. To address this mismatch, we introduce OPD-TV, which regulates the transfer of privileged information based on token-level cross-view compatibility. By comparing the same teacher across both visual states, the proposed method explicitly isolates the additional corrections introduced by changing the teacher view from the standard policy correction in same-view distillation. It then anchors the transfer on the same-view teacher distribution and dynamically scales the transfer strength according to how closely the enriched teacher distribution remains aligned with the student-view teacher distribution. This continuous regulation selectively transfers useful visual refinements while suppressing corrections that are misaligned with the student's visual input. OPD-TV consistently improves upon Video OPD, increasing mIoU by 5.5 and 6.2 percentage points on ActivityNet and QVHighlights, respectively. Our code is available at `https://anonymous.4open.science/r/OPD-TV-Anonymous-6F2A`.

1 INTRODUCTION

On-policy distillation (OPD) provides dense supervision on the states visited by a student model, allowing a teacher to correct the student along its own trajectories (Agarwal et al., 2024; Ross et al., 2011). Privileged on-policy distillation extends this framework by allowing the teacher to access additional information unavailable to the student during training (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016; Penaloza et al., 2026). In Video OPD, such privileged information is instantiated as additional temporal evidence available to the teacher while the student continues to operate under its original visual input. This asymmetry enables the teacher to construct supervision from a richer information state than the student's original visual input (Pinto et al., 2018; Lambert et al., 2018; Gupta et al., 2016). However, privileged information introduces an inherent trade-off: while additional evidence can provide useful corrections, it also changes the teacher's conditional distribution and may introduce target shifts induced by the mismatch between teacher and student visual states (Xue et al., 2026; Yoon et al., 2026; Zhang et al., 2026).

This issue becomes especially important in Video OPD, where privileged information is directly reflected in the teacher view used to construct supervision. However, existing Video OPD avoids this view mismatch by using the same video view for both student rollout and teacher evaluation (Li et al., 2026a), implicitly coupling the student's observation state with the teacher's supervision state and preventing the teacher from exploiting privileged temporal evidence. This design provides an important benefit: the teacher and student are evaluated under identical visual evidence, ensuring that the resulting supervision primarily reflects the policy difference between them. Despite this

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.

advantage, this coupling also constrains the teacher to the student's visual perspective, preventing the teacher from exploiting alternative visual evidence without changing the student trajectory.

To investigate the role of the teacher's visual view, we decouple the view used for student rollout from the view used to construct teacher supervision. We find a supervision transfer gap: enriching the teacher view consistently improves the teacher's grounding accuracy, yet the resulting supervision yields non-monotonic student gains. We reveal that this gap arises because an enriched teacher view introduces not only improved prediction capability but also an additional target shift induced by the difference between teacher and student visual states. When the teacher view becomes overly enriched, this view-induced shift becomes increasingly misaligned with the student's original visual input, limiting the transferability of the supervision.

To address this gap, we introduce OPD-TV, which aims to preserve the benefits of privileged visual evidence while preventing information-induced supervision shifts from being transferred indiscriminately. OPD-TV adaptively regulates the influence of an enriched teacher view based on token-level cross-view compatibility, allowing useful visual corrections to be transferred while reducing mismatched changes between teacher and student supervision states. OPD-TV consistently outperforms standard same-view distillation and direct privileged transfer across temporal grounding benchmarks, demonstrating that effective privileged supervision requires regulating not only the amount of information available to the teacher, but also how that information is transferred to the student.

Our contributions are summarized as follows:

1. We identify the teacher view as an overlooked factor in Video OPD. By decoupling the student view from the teacher view, we show that changing only the teacher's visual information state can substantially alter the resulting supervision.
2. We reveal a supervision transfer gap in privileged Video OPD. While enriched teacher views continuously improve teacher prediction quality, they do not necessarily produce better student learning, exposing the importance of supervision transferability.
3. We introduce OPD-TV, which regulates information-induced supervision shifts through token-level cross-view compatibility. OPD-TV consistently improves upon Video OPD, with mIoU gains of 5.5 and 6.2 percentage points on ActivityNet and QVHighlights, respectively.

2 Preliminaries

2.1 Video On Policy Distillation for Temporal Grounding

Given a video $V = (f_1, \ldots, f_N)$ and a query q, temporal video grounding predicts the corresponding interval $\tau = [t_s, t_e]$ (Gao et al., 2017; Lei et al., 2021). Under a frame budget B, the student processes its original visual input $U_B \subseteq V$, consisting of B frames uniformly sampled from the full video, and autoregressively generates a T-token response $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)$ encoding the predicted boundaries. Video OPD obtains its training states from the current student policy π_θ (Agarwal et al.,

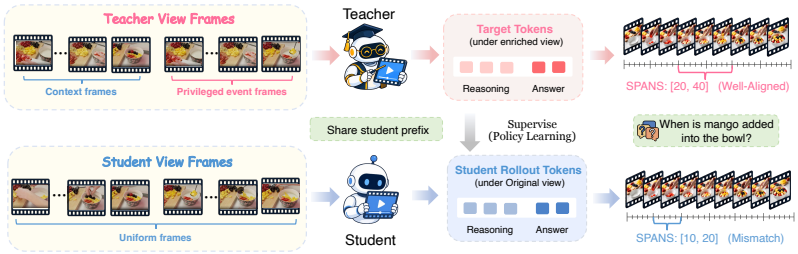

Figure 1: Asymmetric visual information states in Video OPD. The student generates rollout tokens from its original visual input, while the teacher evaluates the corresponding student prefixes using an evidence-enriched view. Changing only the teacher view can therefore alter the supervision target without changing the student trajectory.

2024; Li et al., 2026a). The student first samples $\hat{y} \sim \pi_\theta(\cdot \mid U_B, q)$. At every generation step, the student prefix $\hat{y}_{<t}$ is evaluated by both π_θ and a teacher policy π_ϕ under the same video view, yielding their next token distributions over the vocabulary $\mathcal{V}$:

$$p_t(w) = \pi_\theta(w \mid U_B, q, \hat{y}_{<t}), \qquad q_t(w) = \pi_\phi(w \mid U_B, q, \hat{y}_{<t}). \tag{1}$$

Same view Video OPD minimizes their average divergence D along the sampled response:

$$\mathcal{L}_{\mathrm{VOPD}} = \frac{1}{T} \sum_{t=1}^{T} D(q_t \,\|\, p_t)\,. \tag{2}$$

The rollout determines the prediction states at which supervision is applied, while q_t specifies the teacher's relative preferences over alternative next tokens at each state.

2.2 Visual Information States in Video OPD

Temporal grounding models operate under a finite frame budget B, making their predictions inherently conditioned on a sampled visual view. Video OPD constructs supervision along the student's trajectory, with teacher targets directly shaped by the visual evidence available at each step (Bousselham et al., 2025; Li et al., 2026b). As illustrated in Figure 1, the enriched view $K_m \cup U_{B-m}$ reallocates the shared frame budget: K_m contains m frames sampled from the queried event interval, while U_{B-m} fills the remainder with uniform context frames from the remaining timeline. Exposing a fixed teacher policy π_ϕ to this enriched view defines its conditional distribution over the next token (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023):

$$\widetilde{q}_t(w) = \pi_\phi(w \mid K_m \cup U_{B-m}, q, \hat{y}_{<t}). \tag{3}$$

Conventional Video OPD evaluates both the student and the teacher under the same uniformly sampled original visual input U_B. This symmetric configuration ensures that the resulting target corrects the student based solely on the policy difference between the two policies, as both evaluate the identical prefix under shared visual constraints.

3 Method

We introduce OPD-TV to address the supervision transfer gap: improving the teacher's view does not necessarily produce a better distillation target. OPD-TV regulates the additional correction induced by changing the teacher view. It isolates this correction through paired teacher evaluations in Section 3.1, quantifies cross-view compatibility in Section 3.2, and uses this compatibility to control the correction's contribution to a target anchored to same-view supervision in Section 3.3.

3.1 Isolating the View Correction

Direct privileged distillation forces the student to mimic a target that mixes the teacher's general reasoning capacity with its reliance on newly added visual frames. The student's original visual input may not support the target shifts introduced by the enriched teacher view. We address this mismatch by separating the teacher's baseline capability from the specific token-level adjustments caused by the enriched evidence.

We isolate this visual influence through a paired evaluation. At each generation step t, the student produces a trajectory prefix $\hat{y}_{<t}$ conditioned on its original visual input U_B. We fix the teacher policy π_ϕ and evaluate it on this identical prefix under two distinct visual states: the student's original visual input U_B and the evidence-enriched view $K_m \cup U_{B-m}$:

$$q_t(w) = \pi_\phi(w \mid U_B, q, \hat{y}_{<t}), \quad \widetilde{q}_t(w) = \pi_\phi(w \mid K_m \cup U_{B-m}, q, \hat{y}_{<t}). \tag{4}$$

Evaluating the same network on the identical text query and generation history guarantees that any divergence between q_t and $\widetilde{q}_t$ stems entirely from the altered visual evidence. We quantify this view-induced change as the difference in log-probabilities, $\delta_t(w) = \log \widetilde{q}_t(w) - \log q_t(w)$. This definition explicitly decomposes the enriched-view target relative to the student distribution p_t:

$$\log \widetilde{q}_t(w) - \log p_t(w) = \log q_t(w) - \log p_t(w) + \delta_t(w). \tag{5}$$

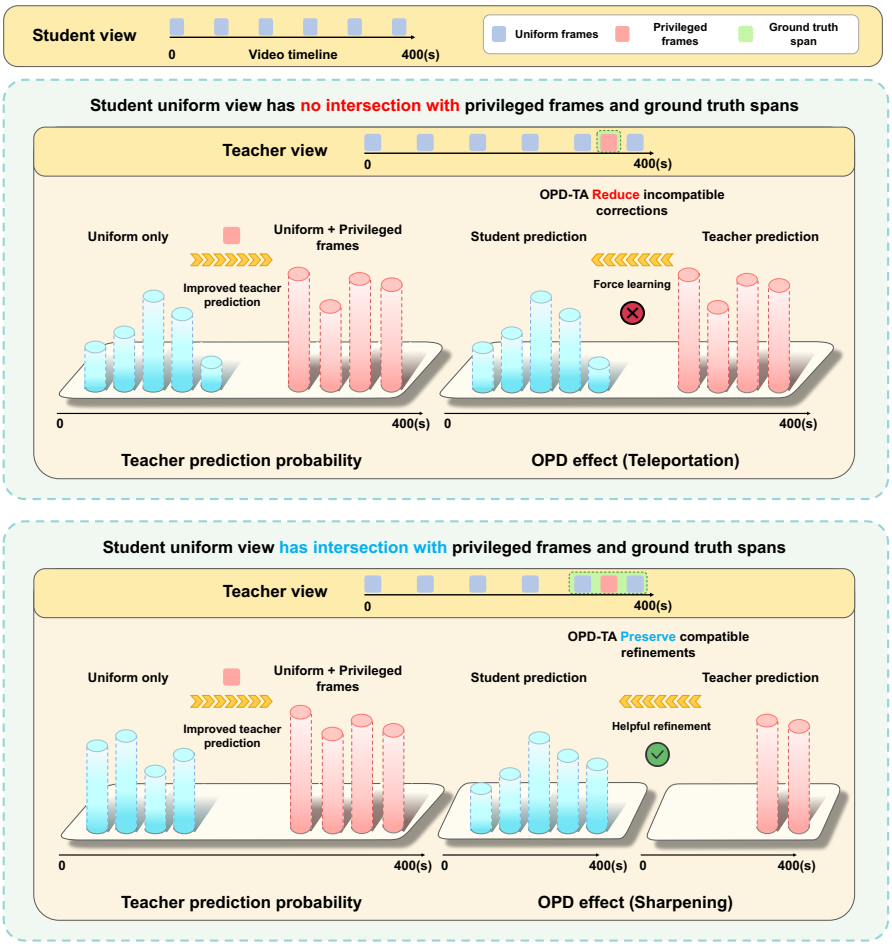

Figure 2: Student coverage of the ground-truth span affects the transferability of privileged supervision. When the student's sampled frames miss the ground-truth span (left), privileged evidence can shift teacher targets toward predictions unsupported by the student's visual input. When the sampled frames overlap the span (right), the additional evidence can refine existing predictions. OPD-TV attenuates incompatible view corrections while preserving compatible refinements.

The first term on the right side provides the standard policy correction, capturing the teacher–student policy difference under the shared visual input. The second term, $\delta_t(w)$, isolates the view correction, tracking exactly how the additional temporal frames shift the teacher's preferences. Decoupling these two signals allows the training process to preserve the baseline policy correction while isolating the specific visual adjustments that require further regulation.

3.2 Quantifying Cross-View Compatibility

The view correction isolated in Section 3.1 captures all visually induced changes, but it does not automatically indicate whether those changes are useful for the student to learn. Additional temporal evidence can either refine preferences among tokens already plausible under the original visual input, or sharply redirect probability mass toward entirely new candidates, as illustrated in Figure 2. To distinguish a transferable refinement from a mismatched target shift, we anchor our measurement on the same-view teacher distribution q_t. We construct a minimal reference set $S_t(\rho)$ containing the most probable tokens that collectively account for a probability mass ρ:

$$k_t(\rho) = \min \left\{ k \in \{1, \dots, |\mathcal{V}|\} : \sum_{i=1}^{k} q_t(w_{(i)}) \geq \rho \right\}, \quad S_t(\rho) = \{w_{(1)}, \dots, w_{(k_t(\rho))}\}. \tag{6}$$

where $w_{(i)}$ denotes the i-th token sorted by decreasing q_t probability, and $\rho \in (0, 1)$ specifies the target probability mass. $S_t(\rho)$ defines a high-probability reference set under the original visual input.

We quantify cross-view compatibility c_t as the total probability mass that the enriched-view teacher $\widetilde{q}_t$ retains within this reference set:

$$c_t = \sum_{w \in S_t(\rho)} \widetilde{q}_t(w). \tag{7}$$

Measuring support retention directly captures the structural impact of the visual evidence. Redistributing probability mass within the reference set leaves compatibility unchanged. Compatibility decreases as more probability mass falls outside the reference set. By characterizing the distributional mismatch between the two target distributions, c_t provides a continuous signal to determine how strongly the view correction should enter the downstream distillation target.

3.3 Regulating View-Correction Transfer

To regulate the transfer of privileged information, we dynamically modulate the influence of the enriched teacher based on its cross-view compatibility c_t. We define a continuous transfer weight $r_t = r_{\min} + (1 - r_{\min})c_t$, where a nonzero minimum $r_{\min} \in (0, 1)$ maintains a nonzero view-correction weight even during strong shifts. The regulated supervision target $\bar{q}_t$ is constructed by balancing the student's adherence to the same-view reference against the enriched preferences via an objective defined on the probability simplex $\Delta(\mathcal{V})$ over the vocabulary, where ν represents a candidate distribution:

$$\bar{q}_t = \arg\min_{\nu \in \Delta(\mathcal{V})} \left\{ (1 - r_t) D_{\mathrm{KL}}(\nu \| q_t) + r_t D_{\mathrm{KL}}(\nu \| \widetilde{q}_t) \right\}. \tag{8}$$

This objective yields a closed-form geometric interpolation that precisely governs how the supervision signal is altered. By normalizing the relative probabilities between any two valid candidates w and v, we expose the structural mechanism of this target in the log-odds space:

$$\log \frac{\bar{q}_t(w)}{\bar{q}_t(v)} - \log \frac{p_t(w)}{p_t(v)} = \log \frac{q_t(w)}{q_t(v)} - \log \frac{p_t(w)}{p_t(v)} + r_t \big[\delta_t(w) - \delta_t(v) \big]. \tag{9}$$

This decomposition shows that the baseline policy correction retains its unit coefficient under regulation. The compatibility-derived scalar r_t gates exclusively the isolated view correction δ_t. By scaling only the visual delta, the target retains more of the view correction at high compatibility and attenuates it at low compatibility.

The student policy is optimized against this dynamically regulated target using a symmetric Jensen–Shannon divergence (JSD) over the training dataset $\mathcal{D}$:

$$\mathcal{L}_{\text{OPD-TV}} = \mathbb{E}_{(V,q,\tau)\sim\mathcal{D}} \mathbb{E}_{\hat{y}\sim\pi_\theta(\cdot|U_B,q)} \left[\frac{1}{T} \sum_{t=1}^{T} \mathrm{JSD}(p_t, \bar{q}_t) \right]. \tag{10}$$

By optimizing this objective, the student generates trajectories strictly from its original visual input while selectively absorbing the privileged visual refinements exposed by the teacher. This continuous regulation mechanism scales view-correction transfer according to cross-view compatibility.

4 Experiments

4.1 Main Results

To evaluate whether privileged visual evidence requires compatibility regulation, we compare OPD-TV against standard fine-tuning (SFT), reinforcement learning (GRPO), and unconditional distillation variants across three temporal grounding benchmarks in Table 1. We observe a hierarchy in grounding performance: while same-view Video OPD provides a robust baseline and Direct Privileged OPD further elevates performance by injecting asymmetric visual evidence, OPD-TV consistently yields the strongest grounding by dynamically regulating this transfer. The absolute percentage-point gain is largest at the strict R@0.7 threshold, where OPD-TV exceeds direct privileged transfer by at least one percentage point across every dataset. Experimental setup and evaluation details are provided in Appendix D.1.

Table 1: Temporal grounding performance across three benchmarks. Direct Privileged OPD improves mIoU over same-view distillation. OPD-TV further improves all reported metrics through compatibility regulation. Both Direct Privileged OPD and OPD-TV use Qwen3-VL-8B-Instruct as the student model.

Method	Charades TimeLens				ActivityNet TimeLens				QVHighlights TimeLens			
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
Proprietary Models												
GPT-4o (OpenAI, 2024)	41.8	60.6	44.5	23.5	40.4	55.2	41.4	25.8	52.1	69.0	54.8	38.5
GPT-5 (OpenAI, 2026)	40.5	59.3	42.0	22.0	42.9	57.4	44.9	30.4	56.8	72.4	60.4	46.4
Gemini-2.0-Flash	46.7	66.4	53.5	27.1	49.3	62.9	54.0	37.7	60.8	76.2	66.4	48.3
Gemini-2.5-Flash (Team, 2025a)	48.6	68.7	56.1	30.6	52.5	66.8	57.5	41.3	64.3	78.2	69.4	55.0
Gemini-2.5-Pro (Team, 2025a)	52.8	74.1	61.1	34.0	58.1	72.3	64.2	47.1	70.4	84.1	75.9	61.1
Open Source Models												
VideoChat-Flash-7B (Li et al., 2025a)	39.7	60.2	37.9	17.8	24.8	35.5	21.8	10.5	32.7	45.2	30.6	16.7
Qwen2.5-VL-7B (Bai et al., 2025)	39.3	59.7	37.8	16.6	31.4	44.1	31.0	16.1	31.6	41.5	27.8	15.2
VideoChat-R1-7B (Li et al., 2025b)	33.7	51.9	30.8	11.7	25.0	35.0	23.9	11.3	21.5	29.3	19.1	9.4
Time-R1-7B (Wang et al., 2025)	36.6	57.9	32.0	16.9	33.1	44.8	31.0	19.0	49.2	65.8	51.5	36.1
Qwen3-VL-32B-Instruct (Team, 2025b)	51.1	72.7	56.0	30.3	51.7	67.0	59.3	41.7	65.5	81.0	69.0	56.7
Qwen3-VL-8B-Instruct (Team, 2025b)	48.3	69.2	53.4	27.5	46.8	62.1	51.2	34.4	59.4	74.2	64.6	49.3
Qwen3-VL-8B (SFT)	40.1	60.7	47.3	20.0	37.1	49.3	39.7	26.0	54.3	66.0	56.0	45.0
Qwen3-VL-8B (GRPO)	50.1	70.7	56.3	30.3	47.8	64.0	52.0	34.0	61.9	78.7	67.7	51.0
Video OPD (Li et al., 2026a)	50.2	72.5	56.1	30.7	49.1	62.8	54.4	37.0	61.4	74.4	64.7	51.9
Direct Privileged OPD (Ours)	51.3	72.0	56.3	31.5	53.4	68.7	59.6	41.7	66.2	81.3	71.9	56.6
OPD-TV (Ours)	**52.0**	**72.7**	**57.2**	**32.5**	**54.6**	**69.5**	**60.3**	**43.2**	**67.6**	**82.1**	**73.5**	**58.3**

4.2 Teacher View Matters in Video OPD

Does changing only the teacher view actually matter? To isolate the effect of the teacher's visual state from model capacity, we conduct a temporal allocation control that holds the underlying policy and the student's original visual input strictly fixed. We vary only how the teacher's frame budget is distributed across the video timeline. Table 2 compares standard same-view supervision against three alternative temporal allocations: an off-target shift concentrated away from the queried event, an annotated-interval view strictly covering the event, and an expanded window embedding the event within its surrounding context. This design separates the general effect of altering the teacher's visual distribution from the specific contribution of query-relevant evidence.

Table 2: Grounding performance under alternative temporal allocations of teacher frames. With the student view and underlying policy held constant, allocating the teacher's visual budget toward query-relevant events through annotated or expanded windows improves grounding performance over uniform or off-target allocations.

Configuration	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**
Qwen3-VL-8B-Instruct	51.5	68.5	56.4	37.1
Same-view OPSD	51.3	66.9	54.4	38.1
Shifted-evidence OPSD	50.9	64.9	54.2	38.4
Annotated-interval OPSD	54.0	**70.6**	56.5	40.4
Expanded-window OPSD	**54.9**	**70.6**	**57.6**	**41.1**

We observe that simply breaking the uniform frame distribution yields no advantage, as the off-target shift behaves similarly to same-view distillation. In contrast, both event-covering views produce distinct grounding improvements. The expanded window performs best overall, indicating that exact annotation boundaries are unnecessary when relevant evidence is provided alongside its temporal context. This difference supports the role of query-relevant visual evidence in target quality. A video teacher can therefore be viewed as a policy–view pair, where changing the temporal view independently alters the supervision target.

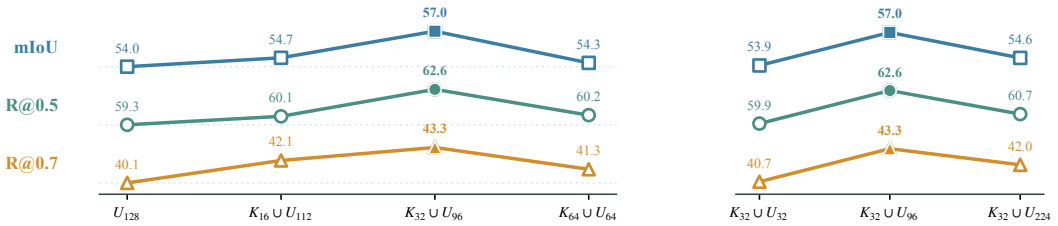

Figure 3: Student performance peaks at intermediate enrichment under direct privileged supervision, whether reallocating a fixed frame budget toward event evidence (left) or adding temporal context with event evidence fixed (right).

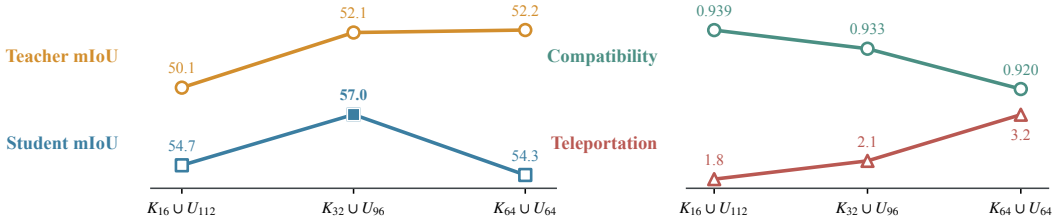

Figure 4: Greater visual enrichment improves teacher accuracy but eventually reduces student performance (left), alongside lower cross-view compatibility and more probability mass outside the same-view reference set (right).

4.3 Better Teacher Views Are Not Always Better Teaching Views

Can we simply choose the view that gives the best teacher? We investigate how the degree of privileged visual evidence affects downstream distillation. The evaluation examines two forms of progressive evidence enrichment in Figure 3: progressively reallocating a fixed frame budget toward the queried event, and expanding the total temporal context while holding the event frames fixed. By tracking both the fixed teacher's own prediction quality and the resulting student performance across these allocations, this design tests whether maximizing the teacher's grounding accuracy automatically provides the optimal supervision target for a student using its original visual input.

We observe a reversal between teacher capability and teaching effectiveness under both forms of progressive evidence enrichment. As event-centric evidence increases, the teacher's own grounding accuracy monotonically improves. The distilled student, however, peaks at a moderate evidence allocation and distinctly degrades under maximum enrichment. This divergence is accompanied by a growing distributional mismatch, as shown in Figure 4. While moderate enrichment sharpens preferences within the same-view reference set, excessive enrichment reduces cross-view compatibility by relocating probability mass away from the support under the original visual input. This suggests that effective teaching depends on the balance between the teacher's predictive advantage and its compatibility with the student's visual state, rather than evidence quantity alone.

4.4 Compatibility Regulation Improves View-Correction Transfer

Does OPD-TV really need token-specific compatibility regulation? To isolate the necessity of continuous, token-level regulation, we compare OPD-TV against alternative strategies that simplify or break how cross-view compatibility determines transfer strength in Table 3. We test whether this local coupling can be replaced by a globally tuned constant (uniform transfer), the correct coefficient distribution randomly assigned across tokens (shuffled transfer), or binary truncation at opposite ends of the compatibility spectrum (Truncate-low and Truncate-high). This design evaluates whether the distillation gain stems from generic target softening, the overall distribution of transfer coefficients, or the precise alignment of each token's correction with support under the student's original visual input.

Table 3: Comparison of view-correction transfer strategies. Replacing continuous, token-level compatibility regulation with binary thresholding, global weighting, or randomized assignments consistently reduces grounding performance. These results support scaling each view correction according to its cross-view compatibility.

Distillation Target	mIoU	R@0.3	R@0.5	R@0.7
Direct privileged	57.0	74.0	62.6	43.3
Uniform transfer strength	57.4	74.2	63.0	43.8
Shuffled transfer assignment	56.8	73.8	62.5	43.2
Truncate-low	55.5	72.9	60.7	43.0
Truncate-high	55.7	72.4	60.8	42.6
OPD-TV	**58.1**	**74.8**	**63.7**	**44.7**

We find that breaking the token-level correspondence between view corrections and compatibility consistently degrades downstream performance. Binary removal limits grounding accuracy at both compatibility extremes, indicating that useful visual refinements are distributed across the spectrum and cannot be isolated by a simple threshold. Furthermore, replacing the adaptive regulation with a globally uniform strength or shuffling valid coefficients across tokens fails to recover the full benefit of the compatibility mechanism. These degradation patterns support the conclusion that the distillation advantage relies directly on matching transfer strength to the local distributional mismatch between the same-view and enriched teacher targets. View correction operates essentially as a dynamically scaled, token-specific adjustment, scaling each refinement according to its compatibility with the student's original visual input.

4.5 Privileged Evidence Improves Direct Distillation

Does privileged evidence add value beyond teacher capacity? To pinpoint the source of distillation gains, we progressively disentangle two dimensions of teacher superiority: policy capacity and visual information state. We first distill an 8B student using a 32B teacher under the same original visual input, measuring the benefit of same-view supervision from a larger teacher. We then hold this 32B teacher policy fixed and enrich its visual state with privileged event evidence. This sequential setup, summarized in Table 4, tests whether the downstream improvements stem entirely from the external teacher's stronger reasoning capabilities, or if the asymmetric temporal view supplies an additional source of useful supervision.

Table 4: Disentangling the sources of distillation gains for an 8B student. Privileged evidence provides additional mIoU gains beyond a larger teacher.

Method	Charades				ActivityNet				QVHighlights			
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
Qwen3-VL-8B-Instruct	48.3	69.2	53.4	27.5	46.8	62.1	51.2	34.4	59.4	74.2	64.6	49.3
+ Strong Teacher	50.5	70.3	54.3	32.0	49.4	66.7	56.7	37.3	61.9	77.7	67.0	51.0
+ Privileged Evidence	51.3	72.0	56.3	31.5	53.4	68.7	59.6	41.7	66.2	81.3	71.9	56.6

We observe that same-view teacher supervision and privileged visual information contribute successive improvements to the distilled student. While the larger same-view teacher reliably enhances the student's baseline grounding capabilities, injecting privileged evidence into that identical fixed policy yields further mIoU gains across all three benchmarks.

4.6 Students Surpassing Their Teachers

Can enriched teacher views help students outperform their teachers? To determine whether the effectiveness of privileged supervision depends on model scale, we vary the capacity and provenance of the teacher policy while holding the 8B student and the evidence-enriched supervision protocol fixed. We compare supervision generated by a larger 32B external model, a matched-capacity

Table 5: Distillation performance across teacher–student capacity configurations. The 8B student matches or exceeds its teacher across all reported benchmarks and metrics.

Evaluated Policy	Charades TimeLens				ActivityNet TimeLens				QVHighlights TimeLens			
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
Qwen3-VL-8B-Instruct	48.3	69.2	53.4	27.5	46.8	62.1	51.2	34.4	59.4	74.2	64.6	49.3
4B-to-8B OPD												
Teacher: Qwen3-VL-4B	47.5	68.0	47.6	26.9	49.1	60.2	54.3	36.5	59.6	74.1	63.2	48.0
Student: Qwen3-VL-8B	**49.8**	**69.4**	**53.9**	**31.5**	**52.4**	**65.2**	**55.6**	**41.5**	**63.2**	**77.8**	**65.5**	**54.0**
8B-to-8B OPSD												
Teacher: Qwen3-VL-8B	46.5	65.7	48.9	27.9	46.5	59.2	51.6	36.2	59.2	73.8	62.8	48.7
Student: Qwen3-VL-8B	**49.2**	**71.7**	**53.2**	**29.2**	**51.6**	**63.9**	**56.3**	**39.5**	**65.1**	**79.8**	**70.8**	**54.7**
32B-to-8B OPD												
Teacher: Qwen3-VL-32B	51.1	72.7	56.0	30.3	51.7	67.0	59.3	41.7	65.5	81.0	69.0	56.7
Student: Qwen3-VL-8B	**52.0**	**72.7**	**57.2**	**32.5**	**54.6**	**69.5**	**60.3**	**43.2**	**67.6**	**82.1**	**73.5**	**58.3**

8B exponential moving average (EMA) teacher for self-distillation, and a smaller 4B model for weak-to-strong distillation in Table 5. We observe that the distilled student matches or exceeds its corresponding teacher on every reported benchmark and metric across all three configurations. The weak-to-strong regime illustrates this finding most clearly: the 4B teacher leveraging privileged event evidence produces supervision that elevates the 8B student beyond the teacher's own grounding performance. This uniform pattern across smaller, matched, and larger policies supports the effectiveness of enriched-view supervision across teacher capacities.

5 Related Work

Temporal Video Grounding. Temporal video grounding maps a language query to the interval containing the corresponding event (Gao et al., 2017; Krishna et al., 2017; Lei et al., 2021). Models perform this prediction from a temporally sampled video view (Lin et al., 2024; Zhang et al., 2024; Li et al., 2025a), and the selected frames determine which parts of the event and its surrounding context are available for grounding (Bai et al., 2023; Wang et al., 2024). Our work examines how this choice of visual evidence affects teacher supervision in temporal grounding.

On-Policy Distillation. OPD constructs dense teacher targets at prefixes generated by the current student (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024; Li et al., 2026c), while OPSD obtains related supervision from a same-source teacher (Furlanello et al., 2018; Zhao et al., 2026). Subsequent work selects informative trajectories (Li et al., 2026d; Jiang et al., 2026; Xu et al., 2026), reweights token-level supervision (Jin et al., 2026), and adapts OPD to temporal video grounding (Li et al., 2026d;a). Video OPD evaluates the teacher on the same video view used by the student, coupling the view that generates the trajectory with the view that constructs the teacher target (Li et al., 2026a). Parallel multimodal approaches enrich teacher supervision with regional crops, recovered visual cues, or counterfactual visual detail (Yuan et al., 2026; Tian et al., 2026; Liu et al., 2026). Building on this use of enriched visual evidence, OPD-TV studies its transferability to a student with a fixed view and regulates the resulting target shifts through cross-view compatibility.

6 Conclusion

This work establishes the teacher's visual information state as an independent determinant of supervision quality in Video On-Policy Distillation. We find that enriching the teacher with privileged temporal evidence consistently improves its absolute predictive accuracy, yet it simultaneously introduces a supervision transfer gap. OPD-TV addresses this distributional mismatch by explicitly isolating the view-induced correction and dynamically regulating its influence through token-level cross-view compatibility. By continuously scaling the transfer strength based on its alignment with the student's visual support, this mechanism allows the student to absorb aligned visual refinements while reducing incompatible target shifts. Consistent improvements across temporal grounding benchmarks support the value of aligning information-induced corrections with the student's visual state, using cross-view compatibility to regulate their contribution to video on-policy distillation.

Ethics Statement

Because on-policy distillation aligns the student with the teacher's preference distributions, the distilled model can inherit social biases and factual errors from the teacher policy. Although cross-view compatibility regulates target shifts induced by asymmetric visual evidence, it measures distributional agreement rather than factual accuracy. It attenuates corrections that lack support under the student's visual input, but does not explicitly evaluate or filter out biased refinements. Consequently, deploying these temporal grounding models in unconstrained environments requires independent safety validation alongside adherence to the original model and dataset licenses.

Reproducibility Statement

We release our complete training and evaluation code at `https://anonymous.4open.science/r/OPD-TV-Anonymous-6F2A`. The formal definitions of the teacher-view decomposition and the compatibility-regulated objective in Section 3 provide the mathematical basis for replicating the method. Because the exact allocation of temporal evidence directly determines the distillation target, the appendices document the precise construction of the paired visual views, the evidence-sensitive training sequence, and its random-sampling control. These appendices further detail the model configurations, optimization hyperparameters, and inference settings required to reproduce the reported temporal grounding improvements.

Use of LLMs

We used LLMs to polish the grammar, phrasing, and readability of the manuscript and to assist with LaTeX and figure/table formatting. The authors take responsibility for the final text, scientific claims, and reported results.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023. doi: 10.48550/ARXIV.2308.12966. URL https://doi.org/10.48550/arXiv.2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: reasoning transfer from llms to vision-language models via on-policy distillation. *CoRR*, abs/2510.23497, 2025. doi: 10.48550/ARXIV.2510.23497. URL https://doi.org/10.48550/arXiv.2510.23497.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075, 2024. doi: 10.48550/ARXIV.2405.21075. URL https://doi.org/10.48550/arXiv.2405.21075.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1602–1611. PMLR, 2018. URL http://proceedings.mlr.press/v80/furlanello18a.html.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5277–5285. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.563. URL https://doi.org/10.1109/ICCV.2017.563.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2827–2836. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.309. URL https://doi.org/10.1109/CVPR.2016.309.

Li Jiang, Haoran Xu, Yichuan Ding, and Amy Zhang. Trajectory-refined distillation. *CoRR*, abs/2606.08432, 2026. doi: 10.48550/ARXIV.2606.08432. URL https://doi.org/10.48550/arXiv.2606.08432.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *CoRR*, abs/2603.07079, 2026. doi: 10.48550/ARXIV.2603.07079. URL https://doi.org/10.48550/arXiv.2603.07079.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 706–715. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.83. URL https://doi.org/10.1109/ICCV.2017.83.

John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8886–8895. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00926. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Lambert_Deep_Learning_Under_CVPR_2018_paper.html.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11846–11858, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/62e0973455fd26eb03e91d5741a4a3bb-Abstract.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. *CoRR*, abs/2602.02994, 2026a. doi: 10.48550/ARXIV.2602.02994. URL https://doi.org/10.48550/arXiv.2602.02994.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22195–22206. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *CoRR*, abs/2606.18974, 2026b. doi: 10.48550/ARXIV.2606.18974. URL https://doi.org/10.48550/arXiv.2606.18974.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *CoRR*, abs/2501.00574, 2025a. doi: 10.48550/ARXIV.2501.00574. URL https://doi.org/10.48550/arXiv.2501.00574.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *CoRR*, abs/2504.06958, 2025b. doi: 10.48550/ARXIV.2504.06958. URL https://doi.org/10.48550/arXiv.2504.06958.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *CoRR*, abs/2604.13016, 2026c. doi: 10.48550/ARXIV.2604.13016. URL https://doi.org/10.48550/arXiv.2604.13016.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, then reweight: Rethinking optimization granularity in on-policy distillation. *CoRR*, abs/2606.02684, 2026d. doi: 10.48550/ARXIV.2606.02684. URL https://doi.org/10.48550/arXiv.2606.02684.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5971–5984. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.342. URL https://doi.org/10.18653/v1/2024.emnlp-main.342.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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. *CoRR*, abs/2605.21924, 2026. doi: 10.48550/ARXIV.2605.21924. URL https://doi.org/10.48550/arXiv.2605.21924.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 8731–8772. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.517. URL https://doi.org/10.18653/v1/2024.findings-acl.517.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.03643.

OpenAI. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. doi: 10.48550/ARXIV.2410.21276. URL https://doi.org/10.48550/arXiv.2410.21276.

OpenAI. Openai GPT-5 system card. *CoRR*, abs/2601.03267, 2026. doi: 10.48550/ARXIV.2601.03267. URL https://doi.org/10.48550/arXiv.2601.03267.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *CoRR*, abs/2602.04942, 2026. doi: 10.48550/ARXIV.2602.04942. URL https://doi.org/10.48550/arXiv.2602.04942.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov (eds.), *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. doi: 10.15607/RSS.2018.XIV.008. URL http://www.roboticsproceedings.org/rss14/p08.html.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 627–635. JMLR.org, 2011. URL http://proceedings.mlr.press/v15/ross11a/ross11a.pdf.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/

arXiv.2402.03300.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 1195–1204, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025a. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025b. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation. *CoRR*, abs/2606.05718, 2026. doi: 10.48550/ARXIV.2606.05718. URL https://doi.org/10.48550/arXiv.2606.05718.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009. doi: 10.1016/J.NEUNET.2009.06.042. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024. doi: 10.48550/ARXIV.2409.12191. URL https://doi.org/10.48550/arXiv.2409.12191.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/7801b29c93b599b8d0c44138596bdeed-Abstract-Conference.html.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13754–13765. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01284. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Number_it_Temporal_Grounding_Videos_like_Flipping_Manga_CVPR_2025_paper.html.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation. *CoRR*, abs/2607.26057, 2026. doi: 10.48550/ARXIV.2607.26057. URL https://doi.org/10.48550/arXiv.2607.26057.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill what the student can see: Fisher-projected on-policy distillation for vision-language models. *CoRR*, abs/2608.01263, 2026. doi: 10.48550/ARXIV.2608.01263. URL https://doi.org/10.48550/arXiv.2608.01263.

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. *CoRR*, abs/2606.00564, 2026. doi: 10.48550/ARXIV.2606.00564. URL https://doi.org/10.48550/arXiv.2606.00564.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *CoRR*, abs/2605.18740, 2026. doi: 10.48550/ARXIV.2605.18740. URL https://doi.org/10.48550/arXiv.2605.18740.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. VAD: attributing visual evidence for target reconstruction in multimodal on-policy distillation. *CoRR*, abs/2607.28590, 2026. doi: 10.48550/ARXIV.2607.28590. URL https://doi.org/10.48550/arXiv.2607.28590.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *CoRR*, abs/2406.16852, 2024. doi: 10.48550/ARXIV.2406.16852. URL https://doi.org/10.48550/arXiv.2406.16852.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *CoRR*, abs/2601.18734, 2026. doi: 10.48550/ARXIV.2601.18734. URL https://doi.org/10.48550/arXiv.2601.18734.

A TEACHER VIEWS CREATE A PREDICTION–TEACHING GAP

Privileged visual evidence can improve a teacher's prediction accuracy without providing equally effective supervision for the student. Because the teacher's token preferences are shaped by temporal frames unavailable to the student, the accuracy of a view-induced correction and its compatibility with the student's visual state are distinct properties of the supervision target.

A.1 CORRECTION QUALITY DETERMINES TRANSFER DIRECTION

Although privileged evidence improves the teacher's aggregate grounding accuracy, this overall gain includes examples where predictions improve, remain unchanged, or move away from the ground truth. To examine how these outcomes affect student learning, we compare transfer strategies separately for each outcome on QVHighlights in Figure 5. This comparison distinguishes the quality of the view correction from the mechanism used to transfer it.

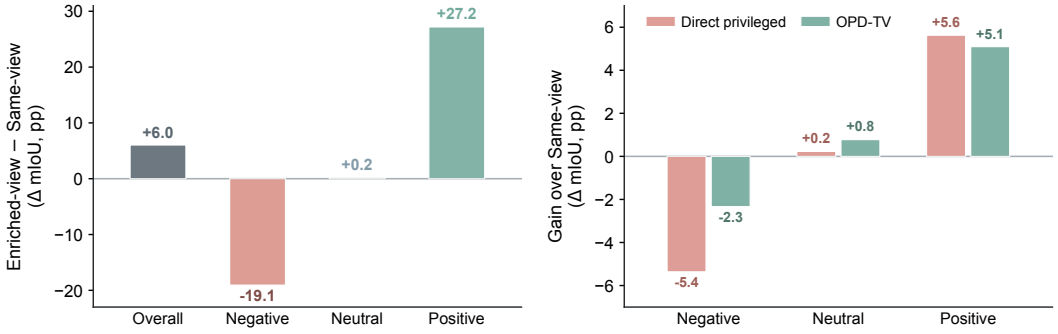

Figure 5: Downstream impact of the teacher's view correction on student learning. When privileged evidence degrades the teacher's prediction, unconditional direct transfer actively harms the student. Regulating this transfer via cross-view compatibility (OPD-TV) suppresses these misaligned target shifts while retaining the gains from positive visual refinements. A controlled 128-frame budget isolates this view-induced effect from baseline training–inference discrepancy.

We observe that the quality of the teacher's correction predicts the direction of the student's performance shift, while the transfer strategy affects its magnitude. When the enriched view improves the teacher's prediction, both direct privileged transfer and OPD-TV improve student performance. When it degrades the teacher's prediction, direct transfer reduces student performance, whereas OPD-TV attenuates this decline through compatibility regulation. The remaining decline shows that compatibility does not identify all harmful corrections. These results support using cross-view compatibility to reduce the transfer of misaligned corrections while retaining gains from useful refinements.

A.2 MORE EVENT EVIDENCE CAN HURT DISTILLATION

To isolate the effect of temporal evidence concentration, we progressively reallocate a fixed teacher frame budget between event-centered frames and uniformly sampled context, as shown in Figure 3. By increasing the proportion of frames strictly covering the queried event while holding the underlying policy and total visual capacity constant, this progressive evidence enrichment specifically isolates the temporal allocation of evidence. This design tests whether continuously maximizing the teacher's exposure to query-relevant information automatically translates into a superior supervision target for the student.

We observe a strict divergence between teacher capability and student learning as event concentration increases. While the heavily enriched teacher continuously improves its own grounding accuracy, the distilled student's performance peaks at a moderate allocation and subsequently degrades. This reversal is accompanied by reduced cross-view compatibility and increased relocation of probability mass away from the support under the original visual input, as shown in Figure 4. Consequently, the most accurate teacher ceases to be the most effective teacher. Teaching quality

Table 6: Downstream distillation performance under matched student and teacher visual budgets with fixed event evidence. The intermediate budget performs best on all reported metrics, showing that additional context does not uniformly improve teaching quality.

Student view	Teacher view	mIoU	R@0.3	R@0.5	R@0.7
U_{64}	$K_{32} \cup U_{32}$	53.9	70.8	59.9	40.7
U_{128}	$K_{32} \cup U_{96}$	**57.0**	**74.0**	**62.6**	**43.3**
U_{256}	$K_{32} \cup U_{224}$	54.6	71.7	60.7	42.0

is governed by balancing the teacher's predictive advantage against its structural compatibility with the student's observation, rather than relying on evidence quantity alone.

A.3 Intermediate Budgets Best Balance Visual Coverage

To determine whether the efficacy of privileged supervision depends on the overall temporal coverage, we scale the total visual budget for both policies simultaneously while preserving a constant absolute amount of event-centered evidence for the teacher. By explicitly varying the surrounding temporal context while locking the core event evidence, this control in Table 6 tests whether effective distillation stems merely from observing the target event, or if it requires a proper balance between focused visual evidence and the broader video timeline.

We find a non-monotonic relationship in overall grounding performance as the temporal context expands. Insufficient context limits performance, suggesting that background frames help the student interpret the teacher's event-specific corrections and localize temporal boundaries. Conversely, further expanding the visual budget beyond the intermediate setting reduces mIoU and recall at all reported thresholds. The intermediate budget best balances these outcomes, supporting the selected configuration as an effective operating point under matched student and teacher frame counts. This pattern reinforces the broader principle that merely maximizing the visual input volume does not automatically yield a better teaching state.

B Compatibility Explains and Regulates View Corrections

Privileged visual evidence alters the teacher's target distribution, but the resulting distributional mismatch varies. Understanding the supervision transfer gap requires distinguishing between targeted preference adjustments within the reference set under the original visual input and shifts that move probability mass outside that candidate space. By analyzing the token-level distributional mismatch induced by these view corrections, we characterize cross-view compatibility as a structural coordinate that distinguishes within-set refinements from cross-set target shifts.

B.1 Compatibility Distinguishes Refinement from Support Shift

To determine whether cross-view compatibility captures the distributional mismatch induced by the view correction rather than merely acting as a proxy for teacher confidence, we examine how the enriched teacher redistributes probability mass relative to the reference under the original visual input. We track two complementary token-level behaviors across different compatibility regimes in Figure 6: the discrete departure of the top prediction from the support set under the original visual input, and the continuous amount of probability mass relocated outside that boundary. This token-level diagnosis tests whether the proposed metric distinguishes a localized preference refinement from a relocation of probability mass outside the reference set.

We find that compatibility distinguishes different degrees of distributional mismatch in the view correction. Under low compatibility, the enriched teacher exhibits a pronounced shift away from the candidate space under the original visual input, placing the majority of its probability mass on tokens outside the reference set. In this regime, the continuous regulation mechanism limits the incompatible shift, sharply restricting the outside mass while retaining the view correction in the target. Conversely, under high compatibility, the enriched teacher primarily refines its preferences

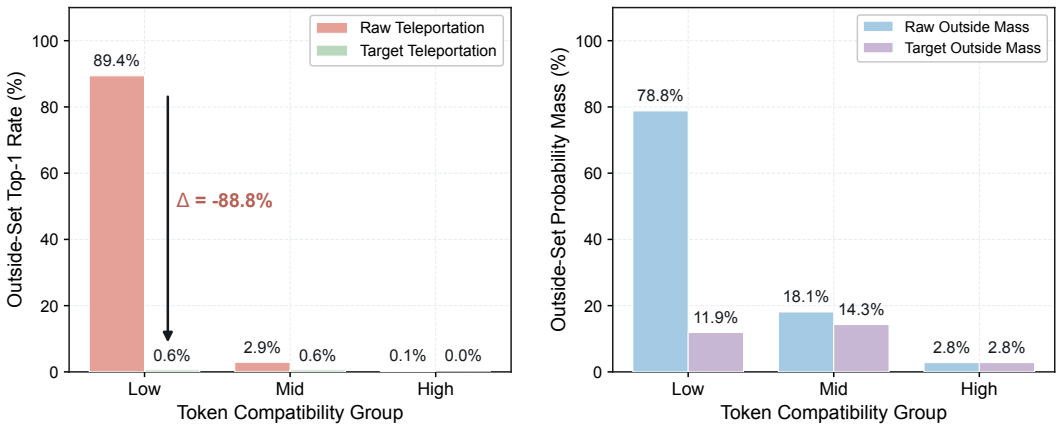

Figure 6: Regulation of distributional mismatch in the teacher's target distribution at temporal decision tokens, measured by top-prediction departures from the same-view reference set and probability mass outside that set. OPD-TV sharply reduces cross-set shifts under low compatibility while largely preserving aligned refinements under high compatibility.

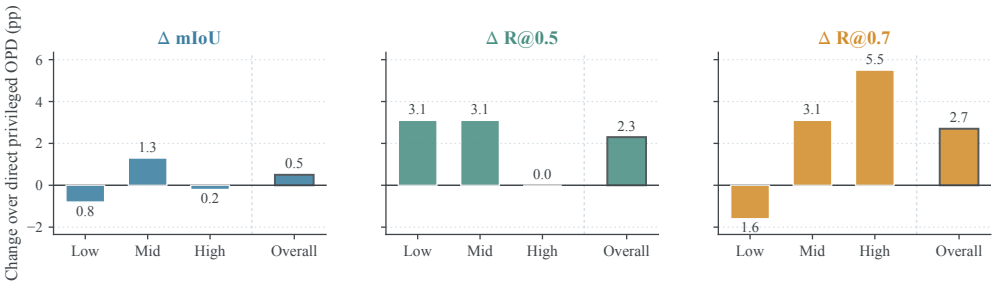

Figure 7: Downstream performance impact of compatibility regulation compared to unconditional direct transfer. The effect of regulation varies across compatibility groups, with mixed outcomes under low compatibility and improved boundary precision under high compatibility. The token-level analysis of distributional mismatch is conducted under a controlled 128-frame budget to explicitly isolate boundary-level mismatches.

among existing reference candidates; the top prediction remains anchored, and the regulation mechanism allows these aligned visual refinements to pass through largely unaltered. This structural correspondence shows how compatibility characterizes cross-set probability relocation, providing a continuous signal for reducing mismatched target shifts while retaining aligned visual refinements.

B.2 Compatibility Predicts Where Regulation Helps

To determine whether continuous regulation selectively affects states with different degrees of distributional mismatch, we compare OPD-TV with direct privileged transfer on the same QVHighlights evaluation examples. We group examples by the minimum compatibility $c_{\min}$ over their temporal boundary tokens: low for $c_{\min} < 0.5$, mid for $0.5 \leq c_{\min} < 0.9$, and high for $c_{\min} \geq 0.9$. The paired comparison in Figure 7 tests how the grounding effect of regulation varies with the strongest boundary-level mismatch within each example.

We find distinct effects across the compatibility spectrum. Mid-compatibility examples improve across all reported metrics, while highly compatible examples show their largest improvement at the strict overlap threshold, consistent with preserving aligned visual refinements. Low-compatibility examples improve at R@0.5 but decline in mIoU and R@0.7, showing that stronger regulation can produce competing effects on grounding accuracy. The overall improvement across metrics supports adapting transfer strength to local compatibility, while the mixed low-compatibility outcomes show that a large target displacement alone does not determine the benefit of regulation.

Table 7: Distillation outcomes under alternative token-wise divergence objectives in same-source OPSD. Symmetric Jensen–Shannon divergence improves all reported metrics over forward and reverse KL, supporting a balanced treatment of teacher and student distributions.

Divergence objective	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**
Forward KL, $D_{\mathrm{KL}}(q_t \| p_t)$	54.1	70.3	57.8	39.8
Reverse KL, $D_{\mathrm{KL}}(p_t \| q_t)$	54.2	71.1	57.8	39.7
JSD, $\alpha = 0.5$	**55.3**	**71.8**	**60.1**	**41.1**

B.3 Binary Removal Cannot Replace Continuous Regulation

To test the contribution of continuous interpolation, we compare OPD-TV with hard removal at opposite ends of the compatibility spectrum in Table 3. Truncate-low replaces the enriched-view target with the same-view teacher target at temporal boundary tokens with $c_t < 0.5$, while Truncate-high applies this replacement at $c_t \geq 0.9$. These tokens remain in the loss and retain same-view policy supervision; only the additional view correction is removed. With the training setup otherwise matched, this comparison tests whether binary selection can recover the benefit of continuously scaling each correction.

We observe that removal at either compatibility extreme degrades grounding relative to continuous regulation. The Truncate-low result suggests that displaced corrections can still contain information that is useful at reduced transfer strength. The Truncate-high result supports retaining aligned refinements instead of removing them from the target. Together, these outcomes support preserving useful guidance across compatibility regimes. Continuous regulation lets mismatched corrections contribute at reduced strength while transferring highly compatible refinements with little restriction.

B.4 JSD Best Balances Teacher and Student Support

To determine how the distributional distance metric affects distillation, we compare forward KL, reverse KL, and symmetric Jensen–Shannon divergence in the same-source OPSD setting. We vary only the divergence D in Equation 2, keeping the teacher update, visual inputs, and training setup fixed, and evaluate all variants at step 125 under the default evaluation configuration. This comparison in Table 7 tests the effect of emphasizing teacher coverage, student-supported modes, or a symmetric balance between the two distributions.

We find that forward and reverse KL yield similar grounding performance, while symmetric JSD improves all reported metrics. The shared advantage across metrics is consistent with balancing teacher coverage and the student's existing preferences when matching the two distributions. These results support symmetric JSD with $\alpha = 0.5$ as the distillation objective in our experiments.

C Generality and Practical Value of OPD-TV

The practical value of privileged distillation depends on its effectiveness across teacher capacities, its preservation of general video understanding, and the computational cost of transferring richer supervision. These dimensions characterize whether compatibility-regulated transfer provides useful grounding improvements beyond a single teacher configuration.

C.1 On-Policy Training Preserves General Video Understanding

To determine whether specialization in temporal grounding compromises broader video understanding, we evaluate the post-trained student on TempCompass (Liu et al., 2024), MVBench (Li et al., 2024), and Video-MME (Fu et al., 2024), following the benchmark selection of Video OPD (Li et al., 2026a). The comparison in Figure 8 includes the base Qwen3-VL-8B model, GRPO, and OPD-TV. All trained students use their original visual input, with privileged frames confined to teacher supervision during training. These evaluations test whether grounding gains are accompanied by a loss of temporal reasoning or broader video comprehension.

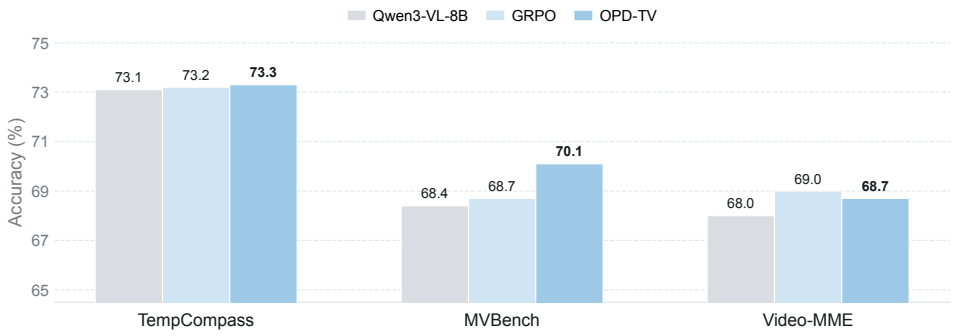

Figure 8: General video understanding following temporal grounding post-training. OPD-TV maintains or improves the base model's scores on TempCompass, MVBench, and Video-MME under the original visual input, showing that grounding gains coexist with preserved performance on these broader evaluations.

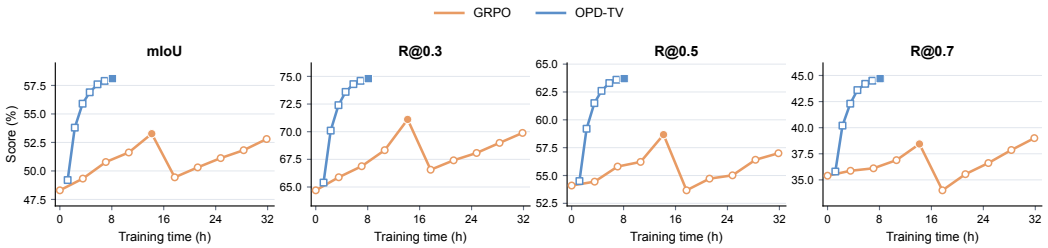

Figure 9: Optimization efficiency of compatibility-regulated distillation versus standard reinforcement learning. OPD-TV reaches higher grounding scores in less training time than GRPO in this comparison, despite the computational overhead of paired teacher evaluations.

We observe that OPD-TV maintains or improves the base model's performance across all three evaluations. It exceeds GRPO on TempCompass and MVBench and remains slightly below it on Video-MME. The preservation of these capabilities shows that the observed grounding gains coexist with general video understanding under the original visual input. Since student trajectories are generated without privileged frames, the additional supervision is learned under the student's own visual condition. This provides a route for transferring task-specific temporal guidance that is consistent with the student's original visual input.

C.2 OPD-TV Retains Favorable Training Efficiency

To evaluate the trade-off between paired teacher evaluations and dense supervision, we compare OPD-TV with GRPO (Shao et al., 2024) using accumulated wall-clock training time. OPD obtains a full next-token distribution from each student trajectory, whereas GRPO uses multiple rollouts to estimate relative advantages from scalar rewards. We evaluate checkpoints every 25 steps under the same default evaluation protocol and compare the resulting grounding trajectories in Figure 9. This comparison tests whether the efficiency of distributional supervision offsets the additional teacher evaluation required by compatibility regulation.

We find that OPD-TV reaches a stronger selected operating point in less training time. OPD-TV achieves 58.1 mIoU and 44.7 R@0.7 after 8.1 hours, compared with 53.3 mIoU and 38.4 R@0.7 for GRPO after 14.2 hours. GRPO also exhibits greater variation across later checkpoints. Despite the additional teacher evaluation, OPD-TV reaches the reported grounding scores in less training time under this setup.

D EXPERIMENTAL PROTOCOL AND PAIRED-VIEW CONSTRUCTION

This section collects the implementation and data-construction details needed to reproduce the preceding experiments. It first specifies the shared model, view, target, and inference configurations, then documents how paired views isolate evidence-sensitive states while holding the policy and query fixed.

D.1 EXPERIMENTAL SETUP

Benchmarks and Evaluation Metrics. We evaluate our approach on three temporal video grounding benchmarks: Charades-STA (Gao et al., 2017), ActivityNet Captions (Krishna et al., 2017), and QVHighlights (Lei et al., 2021). We use the corrected temporal annotations from TimeLens (Wu et al., 2025). We report mean temporal Intersection over Union (mIoU) and Recall@1 at temporal IoU thresholds of 0.3, 0.5, and 0.7. mIoU measures the average overlap between the predicted and annotated intervals, while recall measures the fraction of predictions that reach the specified overlap threshold.

Training Data and Paired Views. We construct paired views from temporally annotated video–query examples in TimeLens-100K (Wu et al., 2025). For each (V, q, τ), the student receives U_{128}, a uniform 128-frame view, while the teacher receives $K_{32} \cup U_{96}$, comprising 32 frames sampled from τ and 96 context frames from the remaining timeline. Both views use the same source-time coordinate system and matched frame budget. We remove unavailable videos and examples that require repeated padding, and limit repeated queries from the same video to preserve scene and temporal diversity. The same frozen Qwen3-VL-8B-Instruct mining policy predicts from both views, isolating the outcome-level change induced by visual information. Let I_U and I_K denote its temporal IoU under the two views and $\Delta I = I_K - I_U$. Selection uses I_U, I_K, and ΔI, followed by duration balancing, to produce a shared 2,400-example online-training sequence. Every target ablation in Table 3 uses this identical evidence-sensitive sequence, isolating the contribution of the distillation target. Complete training and inference configurations are provided in Appendix D.2.

D.2 IMPLEMENTATION DETAILS

Table 8 consolidates the model, visual-input, and target-construction settings used throughout the experiments. All distillation variants use Qwen3-VL-8B-Instruct (Team, 2025b) as the student. External-teacher OPD queries a fixed Qwen3-VL-32B-Instruct model, whereas same-source OPSD forms its teacher as an EMA of the 8B student (Tarvainen & Valpola, 2017; Zhao et al., 2026). Student rollouts use the uniformly sampled original visual input U_{128}; the same-view teacher receives this identical frame sequence, while the evidence-enriched teacher receives the equal-budget view $K_{32} \cup U_{96}$. The paired views therefore differ only in temporal evidence allocation rather than total frame count.

Evaluation retains only the student and applies the same video sampling and decoding configuration to every OPD variant, so the teacher view changes training-time supervision without altering the student's original visual input. Unless otherwise stated, each example contributes one student rollout and the teacher distribution is truncated to its 100 highest-probability tokens before distillation. Cross-view compatibility uses $\rho = 0.9$, OPD-TV sets $r_{\min} = 0.2$, and symmetric JSD uses $\alpha = 0.5$.

D.3 PAIRED VIEWS ISOLATE EVIDENCE-SENSITIVE STATES

We construct the shared training sequence from eligible TimeLens-100K video–query pairs after removing unavailable videos, examples that require repeated frame padding, and excessive repetitions of the same query. The same frozen Qwen3-VL-8B policy then predicts under the original visual input U_{128} and the enriched view $K_{32} \cup U_{96}$. For each pair, we record its IoU I_U under the original visual input, enriched-view IoU I_K, and paired change $\Delta I = I_K - I_U$. These matched outcomes identify states whose grounding behavior responds to visual evidence without changing the policy or query. Duration balancing is applied after this evidence-sensitive selection to retain coverage across short and long events, producing the same 2,400-example sequence for every target ablation. Figure 10 summarizes the complete construction.

Table 8: Model, visual-state, and optimization configurations for external-teacher OPD and same-source OPSD. The paired teacher views use equal frame budgets and differ in temporal evidence allocation, isolating the view-induced correction within each teacher policy.

Configuration	External-teacher OPD	Same-source OPSD
Student backbone	Qwen3-VL-8B-Instruct	
Teacher backbone	Qwen3-VL-32B-Instruct	EMA Qwen3-VL-8B-Instruct
Student rollout view	U_{128}	
Same-view teacher view	U_{128}	
Enriched teacher view	$K_{32} \cup U_{96}$	
Student rollouts per example	1	
Teacher tokens retained	Top 100 by probability	
Distillation divergence	JSD, $\alpha = 0.5$	
Compatibility mass ρ	0.9	
Minimum transfer rate $r_{\min}$	0.2	
EMA update rate	0.001	
Inference sampling rate	2 FPS	
Inference frame cap	768	
Maximum input length	14,336 tokens	

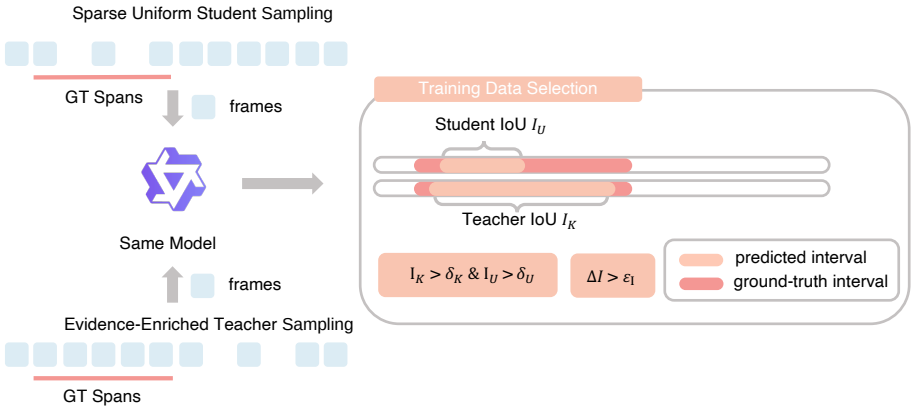

Figure 10: Construction of the evidence-sensitive training sequence. Comparing a shared frozen policy across paired visual states identifies video–query pairs whose grounding predictions respond to temporal evidence; duration balancing preserves coverage across event lengths.

Because both predictions come from a shared policy on the same video–query pair, the selection signal is attributable to the visual information state rather than teacher capacity, model identity, or unmatched examples. The sequence is deliberately enriched for cases in which the two views expose a meaningful supervision difference; it is not intended to approximate random sampling from the full corpus. Reusing it across all target variants prevents data composition from becoming a hidden source of variation, so downstream differences can be traced to how each method constructs the teacher target. Duration balancing adds a complementary safeguard by preventing evidence sensitivity from concentrating the sequence around only one range of event lengths.

D.4 Evidence-Sensitive Selection Drives Generalizable Transfer

To determine whether effective privileged distillation depends on identifying states where the teacher responds to visual evidence, we compare the evidence-sensitive paired-view construction against a matched random-sampling control drawn from the same eligible candidate pool. We hold the dataset size, duration composition, student and teacher policies, visual views, and compatibility-regulated objective fixed, together with the optimization budget and evaluation protocol. By varying only the training-pair selection strategy, this control tests whether exposure to view-induced prediction changes improves supervision transfer. The comparison is reported in Table 9.

Table 9: Downstream generalization under evidence-sensitive and random training-pair selection with matched data size and duration composition. Evidence-sensitive selection improves mIoU on all three benchmarks and R@0.7 on two, with a decline on Charades. Bold marks the best results, including ties.

Benchmark	Selection	mIoU	R@0.3	R@0.5	R@0.7
Charades	Random-2400	51.4	71.0	56.2	**32.9**
Charades	Evidence-sensitive	**52.0**	**72.7**	**57.2**	32.5
ActivityNet	Random-2400	54.3	68.9	58.3	41.9
ActivityNet	Evidence-sensitive	**54.6**	**69.5**	**60.3**	**43.2**
QVHighlights	Random-2400	66.8	**82.1**	69.8	56.0
QVHighlights	Evidence-sensitive	**67.6**	**82.1**	**73.5**	**58.3**
Three-benchmark avg.	Random-2400	57.5	74.0	61.4	43.6
Three-benchmark avg.	Evidence-sensitive	**58.1**	**74.8**	**63.7**	**44.7**

We observe that random sampling yields uneven transfer across benchmarks and lower average grounding performance, failing to consistently retain the strict boundary precision gains achieved by the evidence-sensitive sequence. This pattern suggests that random sampling allocates more updates to examples where additional visual evidence has little effect on the teacher's preferences, reducing the concentration of useful privileged corrections. Evidence-sensitive construction instead selects video–query pairs where the teacher's prediction responds to its visual condition, increasing opportunities to transfer meaningful view corrections. The results support aligning the training sequence with the teacher's cross-view prediction differences to improve the effectiveness and consistency of privileged supervision.

E Properties of the Compatibility-Regulated Target

This section analyzes the regulated target distribution derived from the cross-view interpolation objective. We show that balancing the student's adherence to the same-view and enriched-view teachers yields a closed-form geometric combination that cleanly isolates the view-induced correction and explicitly controls the target's KL departure from the reference under the original visual input.

E.1 Geometric Target Construction

The optimization objective continuously modulates the influence of the enriched-view teacher based on the transfer strength r_t. For a given generation step and fixed positive teacher distributions on a common finite vocabulary, suppressing the token index t, we define the isolated view correction in the log-probability space as $\delta(w) = \log \widetilde{q}(w) - \log q(w)$. The objective is minimized by a target distribution $\bar{q}$ that geometrically interpolates between the two teacher states:

$$\bar{q}(w) = \frac{q(w)\exp(r\delta(w))}{Z(r)} = \frac{q(w)^{1-r}\widetilde{q}(w)^{r}}{Z(r)}, \qquad Z(r) = \sum_{v\in\mathcal{V}} q(v)^{1-r}\widetilde{q}(v)^{r}. \tag{11}$$

We expose the mechanism of this target by rewriting the combined objective for any valid candidate distribution $\nu \in \Delta(\mathcal{V})$:

$$(1-r)D_{\mathrm{KL}}(\nu\|q) + rD_{\mathrm{KL}}(\nu\|\widetilde{q}) = D_{\mathrm{KL}}(\nu\|\bar{q}) - \log Z(r). \tag{12}$$

Because the Kullback–Leibler divergence is non-negative and vanishes only when the distributions perfectly match, $\nu = \bar{q}$ uniquely minimizes the objective. This closed-form solution ensures that the transfer strength r exactly scales the token-level view correction $\delta(w)$. Rather than allowing the enriched teacher's absolute probabilities to overwrite the distillation target, the geometric formulation applies the privileged visual evidence as a regulated logit adjustment, preserving the teacher's baseline policy correction under the shared view.

E.2 Controlling Target Shift via Cross-View Compatibility

Unregulated privileged transfer forces the student to absorb the full enriched-view distribution, potentially introducing severe target shifts that the original visual input cannot ground. We quantify how the regulated target controls this departure by differentiating the log-normalizer $A(r) = \log Z(r)$. The finite sum yields:

$$A'(r) = \mathbb{E}_{\bar{q}}[\delta], \qquad A''(r) = \mathrm{Var}_{\bar{q}}[\delta] \geq 0. \tag{13}$$

Consequently, the KL divergence from the same-view reference is governed directly by the variance of the view correction:

$$\begin{aligned} D_{\mathrm{KL}}(\bar{q}\|q) &= rA'(r) - A(r), \\ \frac{d}{dr} D_{\mathrm{KL}}(\bar{q}\|q) &= r\,\mathrm{Var}_{\bar{q}}[\delta] \geq 0 \qquad (r \geq 0). \end{aligned} \tag{14}$$

This relationship proves that increasing the transfer weight cannot decrease the distillation target's KL departure from the reference under the original visual input. By dynamically coupling the transfer rate r to token-level cross-view compatibility, OPD-TV actively controls this divergence. When privileged evidence triggers aggressive preference shifts outside the same-view reference support—reflected by a drop in compatibility—the reduced interpolation weight limits the target's departure. The target therefore remains anchored to the same-view reference, retaining aligned refinements while reducing incompatible target shifts.