VERIFICATION, NOT CORRECTNESS: LABEL-FREE DISTILLATION OF GROUNDED VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video question answering (VideoQA) models often answer correctly while pointing to the wrong video evidence. Agentic reasoners close much of this *grounding gap* by verifying intermediate claims against observed events, at a steep inference cost. We introduce VERA, which distills a verifier-equipped agentic teacher into a student that emits an answer, typed temporal evidence, and a self-verification score–calibrated to predict the verifier's acceptance of its output–in a single autoregressive pass. A controller trained by preference optimization on the verifier's signal alone raises the teacher's verified yield; a *tilted verification filter* selects and weights traces without answer labels and distills outputs and belief-state representations; and a self-verification head amortizes the verifier itself, enabling a cost-controlled escalation cascade. On NExT-GQA the 4B student reaches 41.0 Acc@GQA, within 1.3 of the teacher's 42.3, at $0.05\times$ its inference cost, and improves on correctness-filtered distillation by $+5.0$ Acc@GQA without answer labels in controller training or trace selection; more than half of this gain comes from a higher localization rate among correct answers. The advantage replicates on a second teacher–verifier pair and on ReXTime and CG-Bench. A supporting analysis expresses the filtered teacher's correct-but-ungrounded mass through measurable verifier primitives and transfers the bound to the student; we report it as a population-level diagnostic with stated limitations.

1 INTRODUCTION

The problem. (Xiao et al., 2024) showed that strong VideoQA models answer many questions correctly while pointing to the *wrong* moments in the video: answer accuracy hides a large *grounding gap*, and a model can be right for reasons it cannot point to.

A running example. Take "*Why does the boy run to the table?*" Two responses select the same correct option and differ only in the evidence they cite. "*To eat cake*," answered from prior expectation with no trigger localized, is correct but *ungrounded*; "*because a cake is brought out*," citing the frames where a detected "cake-placed" event precedes the running, is correct *and* grounded—only the second survives a check of its cited evidence against observed events.

The challenge. Agentic reasoners close much of the gap by changing *how* a model reasons–maintaining an explicit belief state, proposing hypotheses, and *verifying* each against observed events with postcondition and temporal checks; LINGUA (Forouzandeh et al., 2026) is representative (event-driven perception, typed memory, a Belief–Action–Verification loop), and evidential-grounding rewards (Luo et al., 2026) and role-decomposed agents (Liu et al., 2026b) reach for grounding by other routes. But grounding comes at a steep inference-time cost–many model and perception calls per question, branching and backtracking, and hand-set control thresholds–*grounded but slow, and brittle to tune*. Distilling such a teacher into a single-pass student is the obvious remedy, yet the default curation signals–correctness (Zelikman et al., 2022) and self-consistency (Yu et al., 2024)–retain correct-but-ungrounded traces (the "*to eat cake*" answer), so the student learns to be right without pointing. The challenge is to transfer the teacher's *grounding*, not merely its answers, at a fraction of its cost.

Proposed method. Whenever a teacher has a *checkable verifier*—a procedure certifying, per reasoning step, that the step's declared effect is supported by observations–three moves become available. *Deliberate*: let the teacher search, with a *learned* controller steering it toward verifiable trajectories. *Certify*: use the verifier as a label-free curation signal that *shifts trace mass toward grounded rea-*

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.

soning—upweighting the cake-citing trace and filtering out the unsupported one–without treating verifier acceptance as a certificate of grounding. *Amortize*: distill that sub-distribution—its outputs, its belief structure, and *the verifier's judgment itself*–into a single-pass student that emits an answer, typed temporal evidence, and a verifier-acceptance score in one pass. Relative to verifier-filtered distillation (Hosseini et al., 2024; Zhu et al., 2026; Lee & Liang, 2026; Mohammadshirazi et al., 2025), VERA (VErified-trace Reasoning Amortization) differs in three respects we test rather than assert: the curation signal is answer-label-free (an external verifier over *temporal* events), the transfer includes belief-state representations, and the verifier is amortized into the deployed model. These moves address the hand-set thresholds (§2.1), the filters that retain correct-but-ungrounded traces (§2.2–2.3), and the loss of the verifier at deployment (§2.4). **Contributions.**

- *Amortized verified reasoning*: a deliberate–certify–amortize recipe with a controller learned purely from the verifier's signal, a two-dial tilted filter containing correctness filtering, verifier gating, and CoT distillation as boundary cases, typed-evidence and belief-state distillation, and the verifier amortized into a head calibrated to its acceptance decision (§2).
- *Evidence that verification, not filtering per se, transfers grounding*: at matched corpus size and compute, verified filtering beats correctness filtering by $+5.0$ Acc@GQA on NExT-GQA without answer labels ($+3.6$ to $+5.0$ across a second teacher–verifier pair on NExT-GQA and three benchmarks with the LINGUA teacher, at the 4B scale; rising to $+5.5$ at 12B; the second-teacher replication covers only the certify–amortize stages), $+2.8$ of it from a higher localization rate among correct answers; interventional and question-swap controls show the student's evidence use is intervention-sensitive and question-specific.
- *A supporting analysis* (Appendix A): a separation result (Theorem 1)—raw-trace distillation admits no grounding bound independent of the teacher's gap—as the analysis's load-bearing formal claim, plus conditional soundness, coverage, and cascade bounds whose plug-ins are vacuous on NExT-GQA; we use them as a measurement framework, not as deployment guarantees.

Related work (brief; full treatment in Appendix G). Grounded VideoQA systems (Yu et al., 2023; Min et al., 2024; Dang et al., 2025; Dong et al., 2026) supply our external baselines; recent work makes evidence explicit in the reasoning output (Meng et al., 2025; Wang et al., 2026b; 2025; Chen et al., 2025b), conditions localization on question intent (Seo et al., 2026), validates grounding inside an agent (Liu et al., 2025; Montes & Lozano, 2026) or by temporal-logic model checking over video automata (Shah et al., 2026), or rewards it during RL (Luo et al., 2026). Closest to our student is VITED (Lu et al., 2025), which emits answers with evidence chains synthesized to maximize answer likelihood—the correctness-driven signal our controls test against, and a system we compare with directly (Table 1); VideoDistill (Zou et al., 2024) transfers visual content but emits no evidence. Trace curation by consistency (Yu et al., 2024), correctness (Zelikman et al., 2022; Yuan et al., 2023), or a verifier (Hosseini et al., 2024; Lee & Liang, 2026; Zhu et al., 2026; Mohammadshirazi et al., 2025; Lin et al., 2026) precedes us; our loop inherits expert iteration and policy distillation (Anthony et al., 2017; Rusu et al., 2015) with the verifier as the improvement operator. VERA differs in *certifying* evidence with an external, label-free verifier during training, distilling the certified behavior into a single pass, and deploying a calibrated estimate of whether the evidence would survive verification; its cascade relates to selective prediction and cost-aware routing (Geifman & El-Yaniv, 2017; Kadavath et al., 2022; Chen et al., 2023; Liu et al., 2026a) with a signal predicting *verifier* acceptance, not answer confidence.

2 METHOD

VERA separates the *deliberation that discovers grounded answers* from the *model that is deployed* (Figure 1). At training time a controller-driven agentic teacher rolls out several traces per question, a verifier certifies which are internally consistent and grounded, certified traces are distilled into a single-pass student, and *all* traces supervise a self-verification head. At test time the loop is gone: the student answers, localizes evidence, and scores its own groundedness in one decoding pass, escalating to the teacher only when that score is low (worked example: Appendix E).

Setup. An input is $x = (V, q)$ with gold answer $y^\star$ and, when available, gold evidence $g^\star$; grounding is judged by a fixed localization criterion $\mathrm{loc}(\hat{g}, g^\star) \in \{0, 1\}$ (§3). The teacher produces a *trace* $\tau = (b_1, a_1, o_1, v_1, \ldots, b_T, a_T, o_T, v_T, \hat{y}, \hat{g})$: b_t a natural-language *belief state*, a_t a control *action* (script selection, perceptual query, reflect, commit), o_t an *observation*, and $v_t \in \{0, 1\}$ the outcome of *verification*—a postcondition check (does observed evidence satisfy the declared effect?) conjoined

with a temporal-consistency check; COMMIT steps also check answer–evidence consistency. The trace ends in an answer $\hat{y}$ and *typed evidence* $\hat{g}$, tuples (event, $[t_s, t_e]$, role, postcondition); typing is what lets verification be re-run on *student* outputs. With $S(\tau) = \frac{1}{T}\sum_t v_t$ the verified-step fraction and $v_c(\tau)$ the final commit check,

$$\bar{S}(\tau) \;:=\; v_c(\tau)\, S(\tau), \qquad \mathrm{Ver}(\tau) \;=\; \mathbf{1}\big[\bar{S}(\tau) \geq \eta\big] \in \{0,1\}. \tag{1}$$

LINGUA, instantiates b_t via typed memory and v_t via its BAV loop; any teacher exposing the same (b_t, a_t, o_t, v_t) interface with a typed postcondition verifier inherits the same bounds (Corollary 1); we rebuild the pipeline on a second such teacher and sketch a non-video instantiation in Appendix F.

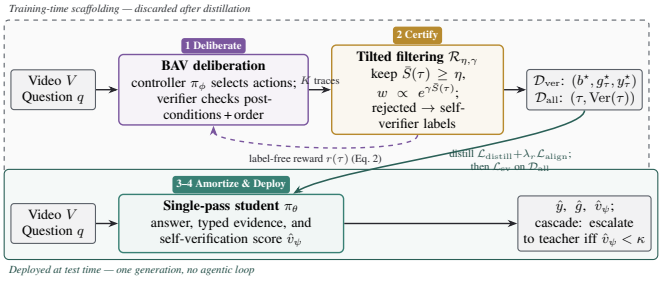

Figure 1: **Overview of VERA.** *Top (training-time scaffolding):* a learned controller π_ϕ drives the teacher's BAV loop, trained with the verifier's signal as a label-free reward (dashed; Eq. (2)); the tilted operator $\mathcal{R}_{\eta,\gamma}$ certifies traces, accepted ones forming $\mathcal{D}_{\mathrm{ver}}$, rejected ones kept in $\mathcal{D}_{\mathrm{all}}$ as self-verifier supervision. *Bottom (deployed):* the student emits answer, typed evidence, and calibrated score $\hat{v}_\psi$ in one pass, escalating inputs with $\hat{v}_\psi < \kappa$.

2.1 DELIBERATE: LEARNING THE CONTROLLER

At step t the controller $\pi_\phi(a_t \mid b_t)$ selects an action; rather than hand-setting these thresholds, we *learn* π_ϕ to maximize an expected verified utility

$$r(\tau) \;=\; \mathrm{Ver}(\tau) \;+\; \alpha\, \mathbf{1}[\hat{y} = y^\star] \;-\; \beta\, c(\tau), \tag{2}$$

where $c(\tau)$ is the deliberation cost and $\alpha, \beta \geq 0$; all main results use $\alpha = 0$, so controller training consumes no answer labels. We optimize π_ϕ by preference learning over trajectories: for each x we sample K traces, rank them by $r(\cdot)$, pair the top-ranked trace with each of the bottom $\lfloor K/2 \rfloor$, and apply a DPO-style objective (Rafailov et al., 2023),

$$\mathcal{L}_{\mathrm{ctrl}}(\phi) \;=\; -\,\mathbb{E}_{(\tau^+,\tau^-)}\Big[\log \sigma\big(\beta_{\mathrm{KL}}\big(\Delta_\phi(\tau^+) - \Delta_\phi(\tau^-)\big)\big)\Big], \quad \Delta_\phi(\tau) = \log \tfrac{\pi_\phi(\tau)}{\pi_{\mathrm{ref}}(\tau)}, \tag{3}$$

with π_{ref} a softened (tempered-softmax) version of the teacher's threshold controller, giving every trajectory positive probability, and per-step log-ratios clipped to bound variance; backbones are *frozen* (details and variants: Appendix B).

2.2 CERTIFY: TILTED VERIFICATION FILTERING

Define the *tilted filtering operator* on trace distributions,

$$\mathcal{R}_{\eta,\gamma}(\pi_{\mathrm{T}})(\tau \mid x) \;\propto\; \pi_{\mathrm{T}}(\tau \mid x)\; \mathbf{1}\big[\bar{S}(\tau) \geq \eta\big]\; e^{\gamma \bar{S}(\tau)}, \tag{4}$$

with two dials: the certification floor η and the tilt γ, which shifts mass toward more thoroughly verified traces ($\gamma = 0$ is rejection sampling). Proposition 1 places correctness-only self-training (Zelikman et al., 2022), hard verifier gating (Hosseini et al., 2024), and unfiltered CoT distillation (Yu et al., 2024) inside this family as boundary cases (as retention rules, not complete algorithms; correctness filtering enters under a substituted trace score, §A.2). η alone determines which inputs are *covered* and can be selected label-free (Proposition 5); γ reweights within the certified region (label-free heuristic: Definition 1). Main results use $\eta = 0.7$, $\gamma = 0.5$ (validation-selected; the verifier-selected variant is within 0.2 Acc@GQA, Table 8). Running the controller-driven teacher with K samples per input and applying $\mathcal{R}_{\eta,\gamma}$ yields

$$\mathcal{D}_{\mathrm{ver}} = \big\{(x,\tau) : \bar{S}(\tau) \geq \eta\big\}, \qquad w(\tau_j \mid x) = \hat{w}(x)\,\omega_j(x), \qquad \omega_j(x) = \frac{e^{\gamma \bar{S}(\tau_j)}}{\sum_{j' \in A(x)} e^{\gamma \bar{S}(\tau_{j'})}}, \qquad \hat{w}(x) \propto \frac{1}{h(\hat{p}_\phi(x))}, \tag{5}$$

where $A(x)$ is the input's accepted-trace set and $h(p) := 1 - (1-p)^K$ estimates the probability that at least one of the K traces survives, from a *cross-fitted* yield estimate $\hat{p}_\phi(x)$ fit on held-out folds (a

same-sample estimate is upward-biased on low-yield inputs; Remark 6); $\hat{w}$ is clipped and normalized to mean one over covered inputs (§A.5). The tilt is normalized *within* each input ($\sum_j \omega_j(x)=1$), so an input's total weight is $\hat{w}(x)$ however many traces it retains, and the inverse-inclusion factor corrects the input shift filtering induces (easy inputs are covered more often) so the objective targets the *original* input distribution: the implemented loss is exactly the input-level (cluster) weighted risk analyzed in Proposition 4. From each retained trace we extract a belief rationale $b^\star$, typed evidence $g^\star_\tau = \hat{g}$, and answer $y^\star_\tau = \hat{y}$; the full pool $\mathcal{D}_{\text{all}} = \{(x, \tau, \text{Ver}(\tau))\}$, rejected traces included, is kept for §2.4. **Verified $\neq$ correct.** Verification certifies internal consistency and evidential support, not answer correctness: a verified trace can carry a wrong answer whose evidence is real. We therefore report the verified-incorrect fraction and compare Ver only (label-free), correctness only (STaR-style), and Ver$\wedge$correct corpora, to test whether verification, not filtering per se, transfers grounding.

2.3 Amortize I: verified belief-state distillation

The student $\pi_\theta(\hat{y}, \hat{g}, b \mid V, q)$ is a single video–language model predicting the answer, typed evidence, and a short rationale in one autoregressive pass ("single-pass" means no agentic loop). Output supervision is weighted maximum likelihood on $\mathcal{D}_{\text{ver}}$:

$$\mathcal{L}_{\text{distill}}(\theta) = \mathbb{E}_{(x,\tau)\sim\mathcal{D}_{\text{ver}}}\, w(\tau)\Big[-\lambda_a \log \pi_\theta(y^\star_\tau \mid x) - \lambda_e \log \pi_\theta(g^\star_\tau \mid x, y^\star_\tau) + \lambda_g\, \mathcal{L}_{\text{ground}}(\hat{g}_\theta, g^\star_\tau) - \lambda_b \log \pi_\theta(b^\star \mid x, y^\star_\tau, g^\star_\tau)\Big], \tag{6}$$

with $\lambda_a = 1$. The student serializes $(y, \hat{g}, b)$ in that order under teacher forcing, so the decoded answer cannot condition on the decoded evidence (the interventional test of §3 masks *video*, not tokens; evidence-first ordering is ablated in Table 8). $\mathcal{L}_{\text{ground}}$ is an IoU-regression penalty on a *differentiable soft span* (expected boundaries under the model's distribution over N_b time bins), so gradients reach the timestamp logits.

Belief-state representation alignment. Output supervision treats belief states as text to paraphrase and leaves the student's *mechanism* unconstrained, so we add a *span-indexed* objective: with the student teacher-forced on the trace's typed evidence, we pool its hidden states at the token positions of span k into $z_{\theta,k}(x)$ and, writing $t(k)$ for the belief step that cited span k and $e(\cdot)$ for a frozen text encoder, minimize

$$\mathcal{L}_{\text{align}}(\theta) \;=\; -\,\mathbb{E}_{(x,\tau)\sim\mathcal{D}_{\text{ver}}}\, \frac{1}{|g^\star_\tau|} \sum_{k=1}^{|g^\star_\tau|} \log \frac{\exp\big(\langle z_{\theta,k}(x),\, e(b_{t(k)})\rangle/\iota\big)}{\sum_{b'\in\mathcal{B}^-\cup\{b_{t(k)}\}} \exp\big(\langle z_{\theta,k}(x),\, e(b')\rangle/\iota\big)}, \tag{7}$$

with M in-batch negatives $\mathcal{B}^-$ drawn from *other* inputs' beliefs and from the same input's *rejected* traces (both sources are ablated); $z_{\theta,k}$ and $e(\cdot)$ are ℓ_2-normalized (Lemma 2). We evaluate the objective independently of QA accuracy using linear probes on $z_{\theta,k}$ for event ordering and postcondition status, compared with Eq. (6) alone (§3).

2.4 Amortize II: self-verification and the escalation cascade

Because evidence is typed, the teacher's verifier *can* be re-run on any output, but that costs fresh perceptual queries over the cited spans, so we amortize *it* too, in two stages. *Stage (i):* a lightweight head $\hat{v}_\psi(x, \hat{y}, \hat{g}) \in [0, 1]$ on the student's final representations is trained on $\mathcal{D}_{\text{all}}$ to predict whether the verifier would certify the trace producing $(\hat{y}, \hat{g})$,

$$\mathcal{L}_{\text{sv}}(\psi) \;=\; \mathbb{E}_{(x,\tau)\sim\mathcal{D}_{\text{all}}}\Big[\text{BCE}\big(\hat{v}_\psi(x, \hat{y}_\tau, \hat{g}_\tau),\, \text{Ver}(\tau)\big)\Big], \tag{8}$$

whose negatives are *real rejected deliberation*—plausible trajectories that failed postcondition or temporal checks. *Stage (ii):* $\mathcal{D}_{\text{all}}$ holds *teacher* outputs whereas $\hat{v}_\psi$ is deployed on *student* outputs, so after distillation we run the verifier offline on the student's typed outputs over a held-out fitting split, fine-tune $\hat{v}_\psi$ on these labels, and temperature-scale it on a disjoint calibration split from the same student-output distribution (Guo et al., 2017)—the distribution against which the calibration error of Proposition 6 is defined. $\hat{v}_\psi$ is thus calibrated to the *verifier's acceptance decision* on student outputs, not directly to gold groundedness; Table 17 reports its calibration and ranking against the joint correct-and-localized event as well.

Selective prediction and cascade. At deployment, $\hat{v}_\psi$ enables selective prediction (Geifman & El-Yaniv, 2017) and an *escalation cascade* that routes inputs with $\hat{v}_\psi < \kappa$ to the agentic teacher (cf.

Chen et al., 2023); Proposition 6 bounds the cascade's grounding gap and Proposition 7 turns it into an operating-point rule. Typing also admits a head-free *verify-then-escalate* control (run the actual verifier on every output, escalate on failure), which pays the verifier's perception cost on every input; comparing the two at matched escalation rates prices what amortization buys (Algorithm 1). **Analysis, in brief** (Appendix A). Theorem 1 shows raw-trace distillation admits no grounding bound independent of the teacher's gap; conditional soundness, coverage, and cascade bounds (Propositions 2–6) are population-level diagnostics whose plug-ins are vacuous on NExT-GQA (summary: §A).

3 EXPERIMENTS

We ask: **Q1** does VERA retain the teacher's *grounded* accuracy at a fraction of the cost, and does it replicate across teachers and benchmarks; **Q2** is *verification*, not filtering per se, the active ingredient; **Q3** does the student use the localized evidence, question-specifically, under randomized controls; **Q4** does alignment transfer belief-state *structure*; **Q5** is the self-verifier calibrated, and does the cascade trace a useful cost–grounding frontier?

Datasets. Grounded-accuracy claims rest on *three* benchmarks with a joint answer-plus-span metric: NExT-GQA (Xiao et al., 2024) (primary), ReXTime (Chen et al., 2024) (causal reasoning across separated events), and CG-Bench (Chen et al., 2025a) (27-minute videos on average). NExT-QA (Xiao et al., 2021), IntentQA (Li et al., 2023), and EgoSchema (Mangalam et al., 2023) lack spans and evaluate answer accuracy. Column-by-column training-split provenance for Table 1 is in Appendix C; Ego4D-NLQ (Grauman et al., 2022) evaluates localization transfer. The student sees 32 frames per query on NExT-GQA and ReXTime and 128 on CG-Bench ($\approx$13 s apart), the teacher's budget matched per benchmark (Table 11).

Students and teachers. Students are Gemma 3 instruction-tuned checkpoints at 1B, 4B (default), and 12B (Team et al., 2025); the 4B and 12B are natively multimodal, and the text-only 1B is paired with the same frozen vision tower and a trained projector (Appendix B). The primary teacher is a verifier-equipped agentic reasoner (event-driven perception, typed memory, and a Belief–Action–Verification loop) run with our implementation plus the learned controller of §2.1; to test whether the curation signal is tied to one implementation, we also use the public agentic teacher VideoMind (Liu et al., 2026b), whose verifier role supplies v_t in place of ours (teacher and verifier vary together; η, γ re-selected label-free).

Baselines (three families). *(a) Published grounded VideoQA systems*: SeViLA (Yu et al., 2023), TimeChat (Ren et al., 2024), MoReVQA (Min et al., 2024), MUPA (Dang et al., 2025), LeAdQA (Dong et al., 2026), the agentic VideoMind-7B (Liu et al., 2026b), the single-pass evidence-chain model VITED (Lu et al., 2025) (the closest published counterpart of our student), Grounded-VideoLLM (Wang et al., 2024), and an open video–language model prompted zero-shot for an answer plus timestamps (Qwen2.5-VL-7B-Instruct, (Bai et al., 2025)), anchoring what an off-the-shelf VLM achieves; prompts and decoding for every re-evaluated (‡) baseline are in Appendix C. *(b) Distillation-signal controls*: identical student, pipeline, and trace budget, varying only the curation signal—zero-shot, raw-CoT distillation (typed evidence *stripped*), correctness-filtered (STaR-style) (Zelikman et al., 2022; Yuan et al., 2023), self-consistency-filtered (Yu et al., 2024; Wang et al., 2023), all-trace (typed evidence kept, full $K{=}8$ pool), verified (ours), and verified$\wedge$correct, all matched on retained examples, covered inputs, target tokens, and optimization budget (Appendix B). *(c) Supervision and deployment controls*: a *gold-evidence supervised* student (ceiling), *test-time verifier reranking* (best-of-K teacher rollouts), a *V-STaR-style* control (best-of-N student samples reranked by a trained verifier; (Hosseini et al., 2024)), *verify-then-escalate*, max-probability and random routing, and the agentic teacher.

Metrics and protocol. Answer accuracy (Acc@QA) and the joint metric Acc@GQA, computed with the *official NExT-GQA evaluator* (Xiao et al., 2024), which credits a question only when the answer is correct *and* $IoP \geq 0.5$; mIoP and mIoU (reported in Tables 2 and 3); and the stricter Acc@GQA^{IoU} ($IoU \geq 0.5$), reported separately and never against published numbers. On the NExT-GQA population we also report the absolute grounding gap $G = \mathrm{Acc@QA} - \mathrm{Acc@GQA}$ (the mass the analysis bounds) and the conditional localization rate $P(\mathrm{loc} \mid \mathrm{corr}) = \mathrm{Acc@GQA}/\mathrm{Acc@QA}$; since $\mathrm{Acc@GQA} = \mathrm{Acc@QA} \times P(\mathrm{loc} \mid \mathrm{corr})$, any grounded-accuracy difference decomposes exactly into an answer part and a localization part. Inference cost is single-query latency and FLOPs relative to the teacher under a matched protocol (details, batching caveats,

and absolute costs: Appendix C); the *amortization break-even* is the query volume at which the offline cost is repaid. The interventional *GFS* ablates predicted spans and measures answer flips (sensitivity) and perturbs irrelevant spans and measures stability (invariance), against a matched-length randomized-ablation control and a *question-swap* control. Calibration of $\hat{v}_\psi$ is binned ECE and $\hat{\xi} = \mathbb{E}|\hat{v}_\psi - Y|$ against the re-run verifier's decision Y; discrimination against gold groundedness is reported separately (Table 17). Results are mean±std over three seeds; pairwise comparisons among our-pipeline systems use a paired bootstrap over test questions (10^4 resamples). *Label use:* the controller reward (α=0), filtering, distillation targets, and self-verifier labels consume no gold answers or spans; annotations enter only validation-split model selection (§2.2), a 1,000-clip instrumentation subset for measuring the analysis, and evaluation.

Table 1: **Main results.** Acc@GQA is the official NExT-GQA joint metric ($IoP \geq 0.5$); Acc@QA is answer accuracy on the same NExT-GQA population; $P(\mathrm{loc}\,|\,\mathrm{corr}) = \mathrm{Acc@GQA}/\mathrm{Acc@QA}$ (%). Our-pipeline results are mean±std over three seeds; by paired bootstrap, VERA vs. each of the five controls above it is significant at $p < 0.01$ on Acc@GQA, while VERA+cascade vs. verify-then-escalate and vs. the gold-span ceiling are *not* ($p > 0.3$). VideoMind-7B is its released 32-frame setting. †Trained with gold spans. ‡Not reported in the cited paper under this protocol; measured by us with the official evaluator and matched cost protocol (SeViLA's NExT-GQA entries are from (Xiao et al., 2024)). §At 20% escalation; cost $c_\mathrm{s} + c_\mathrm{ver} + 0.2\,c_\mathrm{T}$.

	NExT-QA	NExT-GQA			EgoSchema	IntentQA	Cost vs. teacher↓	
Method	Acc	Acc@QA	Acc@GQA	$P(\mathrm{loc}\,\|\,\mathrm{corr})$	Acc	Acc	Latency	FLOPs
Published grounded VideoQA systems and off-the-shelf references								
SeViLA	73.8	68.1	16.6	24.4	22.7	71.0‡	2.6×‡	2.4×‡
TimeChat	73.5‡	70.2‡	19.4‡	27.6	60.3‡	72.8‡	2.1×‡	1.9×‡
MoReVQA	69.2	67.9‡	31.0	45.7	51.7	72.2‡	2.8×‡	2.5×‡
MUPA-7B	74.8	74.1‡	30.3	40.9	61.8‡	72.5‡	1.9×‡	1.8×‡
LeAdQA	75.1‡	74.6‡	32.0‡	42.9	62.1‡	72.9‡	2.3×‡	2.1×‡
VideoMind-7B (agentic)	72.4	71.8	28.2	39.3	58.6	70.9	1.4×‡	1.3×‡
VITED (single-pass)	74.2	73.5	31.6	43.0	60.8	71.5	0.06×‡	0.06×‡
Grounded-VideoLLM	71.9	70.1	26.7	38.1	55.0	68.4	0.05×‡	0.05×‡
Qwen2.5-VL-7B (zero-shot)‡	77.6	75.9	24.8	32.7	63.1	72.0	0.08×	0.08×
Distillation-signal controls (identical student and pipeline)								
Zero-shot student	72.8±0.3	68.9±0.4	28.5±0.4	41.4	55.4±0.5	67.9±0.4	0.04×	0.05×
CoT distillation	79.1±0.3	74.4±0.3	34.2±0.4	46.0	61.3±0.4	72.1±0.3	0.05×	0.05×
Correctness-filtered (STaR)	80.2±0.3	76.3±0.3	36.0±0.3	47.2	62.0±0.3	72.6±0.2	0.05×	0.05×
Self-consistency-filtered	80.4±0.2	76.8±0.3	36.4±0.4	47.4	62.2±0.3	72.7±0.3	0.05×	0.05×
All-trace distillation	80.0±0.3	77.4±0.3	37.1±0.3	47.9	62.4±0.4	72.9±0.3	0.05×	0.05×
VERA (ours)	**81.2**±0.2	81.0±0.2	**41.0**±0.3	**50.6**	**64.8**±0.3	**74.1**±0.2	0.05×	0.05×
Verified ∧ correct (label-dependent)	81.6±0.2	81.5±0.2	41.4±0.3	50.8	64.9±0.4	74.2±0.3	0.05×	0.05×
VERA + cascade (20% escalation)	**81.9**±0.2	81.9±0.2	**41.8**±0.2	**51.0**	**65.4**±0.3	**74.4**±0.2	0.25×	0.25×
Supervision and test-time controls								
Gold-span student (ceiling)†	81.6±0.2	81.9±0.2	42.0±0.3	51.3	63.9±0.3	73.6±0.3	0.05×	0.05×
Verifier reranking (best-of-K=8 teacher)	82.6±0.2	82.6±0.2	42.8±0.2	51.8	66.7±0.3	75.0±0.2	8.0×	8.0×
V-STaR-style rerank (best-of-N=8 student)	81.8±0.2	81.6±0.2	41.5±0.3	50.9	65.0±0.3	74.3±0.2	0.42×	0.40×
Verify-then-escalate (head-free)§	82.0±0.2	82.1±0.2	41.9±0.3	51.0	65.6±0.3	74.5±0.2	0.35×	0.35×
Teacher: LINGUA (agentic)	82.4	82.2	42.3	51.5	66.2	74.8	1.0×	1.0×

Grounded accuracy at $0.05\times$ the cost, and it replicates (Q1). The 4B student reaches 41.0 Acc@GQA, within 1.3 of the agentic teacher's 42.3, at $0.05\times$ its latency and FLOPs (Table 1), within 1.0 of the gold-span ceiling (42.0; the cascade closes this to 0.2, $p > 0.3$) and above every published system we re-evaluated under the official protocol; on NExT-QA, EgoSchema (zero-shot transfer), and IntentQA, which have no spans, answer accuracy improves by +2.1, +3.5, and +2.0 over CoT distillation. The 20%-escalation cascade recovers 41.8 at $0.25\times$, matching head-free verify-then-escalate (41.9, $p > 0.3$) at $0.35\times$; test-time reranking is more accurate (42.8) but costs $8.0\times$. Amortization against the teacher is paid off after $Q^\star \approx 2.8 \times 10^5$ queries; against verify-then-escalate—which matches the cascade's accuracy at zero offline cost—the cascade's $0.10\times$ per-query saving repays the offline bill only above $\approx 2.7 \times 10^6$ queries, so below that volume verify-then-escalate is the rational choice (Appendix C). Table 2 replicates the effect: with VideoMind as teacher *and* verifier, and on ReXTime and CG-Bench, the advantage of verified over correctness filtering stays between +3.6 and +5.0 Acc@GQA, and the student stays within 1.3 of its teacher in every setting.

Verification is the active ingredient (Q2). The decisive comparison is against *correctness filtering*, the field's default curation signal. Holding the student, pipeline, and trace budget fixed and varying only the selection signal (Table 3), verified filtering beats correctness filtering by +5.0 Acc@GQA, +4.4 Acc@GQA^{IoU}, and +4.7 mIoP despite *not* using answer labels, although correctness filtering

Table 2: **Replication across teachers and benchmarks.** Top: the pipeline rebuilt with VideoMind (Liu et al., 2026b) as teacher and verifier (v_t from its verifier role; η, γ re-selected label-free; budgets matched; no controller learned, so only the certify–amortize stages are replicated). The "Teacher" cell is a single greedy agentic pass with the verifier role active at the matched 96-frame budget; the vanilla row is the released 32-frame setting without verifier re-query, hence the 10.3-point difference with the planner untouched (Appendix C). Bottom: three grounded benchmarks with the LINGUA teacher (single greedy pass), each trained on its own split; CG-Bench uses its released clue-grounded protocol. Cells: joint / localization; Δ = verified − correctness-filtered. Replication cells are single runs; per-Δ paired-bootstrap 95% CIs over test questions are in Table 12.

Teacher / benchmark	Metrics	Zero-shot	CoT	Corr.-filt.	VERA	Teacher	Δ
VideoMind-7B / NExT-GQA	Acc@GQA / mIoP	28.5 / 35.2	32.4 / 40.0	33.8 / 41.1	**37.9 / 45.4**	38.5 / 46.3	+4.1 / +4.3
VideoMind-7B / NExT-GQA	GFS	0.41	0.48	0.52	**0.61**	0.63	+0.09
VideoMind-7B (vanilla, ref.)	Acc@GQA / mIoP	–	–	–	–	28.2 / 39.0	–
LINGUA / NExT-GQA (Xiao et al., 2024)	Acc@GQA / mIoP	28.5 / 35.2	34.2 / 38.4	36.0 / 40.2	**41.0 / 44.9**	42.3 / 45.7	+5.0 / +4.7
LINGUA / ReXTime (Chen et al., 2024)	Acc@GQA / mIoP	22.1 / 30.4	26.8 / 33.5	28.4 / 35.1	**32.6 / 39.2**	33.9 / 40.3	+4.2 / +4.1
LINGUA / CG-Bench (Chen et al., 2025a)	acc.@IoU / mIoU	12.4 / 11.3	15.1 / 13.6	16.2 / 14.4	**19.8 / 17.5**	21.0 / 18.3	+3.6 / +3.1

Table 3: **Effect of the trace-selection signal (Q2)** on NExT-GQA; student, pipeline, retained-example count, covered inputs, target-token and optimization budgets identical across rows. $P(\mathrm{loc}\,|\,\mathrm{corr}) = \mathrm{Acc@GQA}/\mathrm{Acc@QA}$ (%); GFS as in Table 4; %ver.-inc. is the fraction of the retained corpus that is verified ∧ incorrect (verifier re-run on the underlying typed traces; for raw CoT, before evidence is stripped).

Corpus	Acc@QA	Acc@GQA	Acc@GQA^{IoU}	mIoP	$P(\mathrm{loc}\,\|\,\mathrm{corr})$	GFS	%ver.-inc.
Raw CoT (unfiltered)	74.4	34.2	26.5	38.4	46.0	0.53	3.1
Correctness (STaR)	76.3	36.0	28.0	40.2	47.2	0.55	0
Self-consistency	76.8	36.4	28.4	40.6	47.4	0.56	2.4
All traces	77.4	37.1	29.0	41.5	47.9	0.59	3.1
Ver only (ours, label-free)	81.0	41.0	32.4	44.9	50.6	0.66	6.8
Ver∧correct	81.5	41.4	32.8	45.2	50.8	0.67	0
Δ (Ver only − correctness)	+4.7	+5.0	+4.4	+4.7	+3.4	+0.11	–

Table 4: **Interventional grounding (Q3)** on NExT-GQA. Δ_{rand} is targeted-minus-random sensitivity; Q-swap is question-swap agreement (chance 0.5).

Method	Sens.↑	Δ_{rand} ↑	Inv.↑	GFS↑	Q-swap↑
Zero-shot student	0.31	0.09	0.62	0.41	0.54
CoT distillation	0.42	0.15	0.70	0.53	0.61
VERA (ours)	0.61	0.34	0.71	0.66	0.79
Teacher (agentic)	0.64	0.37	0.73	0.68	0.83

Table 5: **Cascade frontier (Q5)** on NExT-GQA vs. escalation rate $1 - r_\kappa$, with routing controls at matched rates. Costs follow $c_{\mathrm{s}} + (1 - r_\kappa)c_{\mathrm{T}}$ (so the 100% column exceeds 1.0×); verify-then-escalate also pays $c_{\mathrm{ver}} = 0.10\,c_{\mathrm{T}}$ on every input. ECE of calibrated $\hat{v}_\psi$: 0.045.

Escalation rate	0%	10%	20%	100%
Acc@GQA, $\hat{v}_\psi$-routing (ours)	41.0	41.5	41.8	42.3
Acc@GQA, verify-then-escalate	41.0	41.6	41.9	42.3
Acc@GQA, max-prob routing	41.0	41.2	41.4	42.3
Acc@GQA, random routing	41.0	41.1	41.3	42.3
Cost, $\hat{v}_\psi$-routing	0.05×	0.15×	0.25×	1.05×
Cost, verify-then-escalate	0.15×	0.25×	0.35×	1.15×

zeroes the verified-incorrect rate by construction. Because $\mathrm{Acc@GQA} = \mathrm{Acc@QA} \times P(\mathrm{loc}\,|\,\mathrm{corr})$, the +5.0 decomposes exactly into +2.2 from answering more questions (+4.7 Acc@QA at correctness filtering's localization rate of 47.2%) and +2.8 from localizing a larger share of correct answers (47.2% → 50.6%). Adding correctness to verification buys only +0.4, and the two evidence-bearing unfiltered corpora (all-trace vs. raw CoT) show typed evidence alone contributes +2.9: selection by verification, not evidence format, carries most of the gain. Acceptance-conditioned statistics linking verification to the metric (56.0% vs. 41.2% localization among accepted- vs. rejected-and-correct traces) are in Table 14; the accepted-but-unlocalized remainder is why verification is a curation signal, not a certificate.

Where the grounded gain comes from. Two regularities fix how to read Table 1: $P(\mathrm{loc}\,|\,\mathrm{corr})$ rises monotonically with the grounding signal (41.4 zero-shot → 50.6 VERA → 51.5 teacher) while the gap $G = \mathrm{Acc@QA} - \mathrm{Acc@GQA}$ lies in 39.8–40.4 for *every* in-pipeline system, and the advantage is concentrated on the span-annotated questions (the five controls lose 2.6–4.7 answer-accuracy points from NExT-QA to the NExT-GQA subset; VERA loses none). The per-question analysis behind both—including the 31% of questions no in-pipeline system localizes—is in Appendix C.

Interventional grounding with randomized controls (Q3). Correlational grounding metrics can be gamed, so we test grounding *causally*, a video port of comprehensiveness/sufficiency evaluation

(DeYoung et al., 2020; Jacovi & Goldberg, 2020) (Table 4). VERA nearly closes the gap to the teacher (GFS 0.66 vs. 0.68) and more than doubles CoT distillation's targeted-minus-random sensitivity gap (Δ_{rand} 0.34 vs. 0.15), so the flips are not an artifact of off-distribution masking. The question-swap control separates question-specific grounding from generic salience: on videos carrying several questions with disjoint gold spans, VERA's predicted spans move to their own question's region 79% of the time (chance 50%; CoT distillation 61%), which no question-independent detector can achieve.

Table 6: **Belief-state probing (Q4)**: linear-probe accuracy on $z_{\theta,k}$ from independent gold annotations. Chance is 50% (order) and 33% (postcondition status). The two control rows address teacher-forced textual shortcuts in Eq. (7).

Representations from	Event order↑	Postcond.↑
CoT-distilled student	61.2	44.5
VERA without $\mathcal{L}_{\text{align}}$	64.8	48.9
VERA (full)	74.5	60.7
probed at pre-evidence positions	71.8	57.4
trained with de-lexicalized evidence	72.6	58.8
Teacher belief encodings $e(b_t)$ (reference)	80.3	68.2

Table 7: **Generalization.** Top: NExT-QA accuracy and NExT-GQA Acc@GQA by student scale, for VERA and the correctness-filtered control trained identically at each scale (Δ = VERA − corr.-filt.). Bottom: zero-shot transfer (n/a = undefined). Ego4D-NLQ: R@1@IoU0.5 on NLQ v2 validation queries, 32 frames, official evaluator, no NLQ training (not comparable to trained NLQ systems). Scale rows are means over three seeds; per-seed values and stds are in Table 13.

Student scale	NExT-QA Acc	Acc@GQA (VERA)	Acc@GQA (corr.-filt.)	Δ
1B	78.4	37.2	33.4	+3.8
4B	81.2	41.0	36.0	+5.0
12B	82.1	42.4	36.9	+5.5

Transfer (train→test)	Acc	Grounding
NExT→IntentQA	71.5	n/a
IntentQA→NExT	79.2	38.4
NExT→Ego4D-NLQ	n/a	17.6

Representation probing (Q4). Frozen linear probes on $z_{\theta,k}$ classify the temporal order of gold event pairs and the postcondition status of cited evidence from *independent gold annotations*, not teacher text (Table 6). The measured probe results show that $\mathcal{L}_{\text{align}}$ raises event-order probing from 64.8 to 74.5 and postcondition-status probing from 48.9 to 60.7 at matched QA accuracy, approaching the teacher's own belief encodings (80.3/68.2). Since aligned positions carry teacher-forced evidence tokens, Eq. (7) could be satisfied by textual co-occurrence; probing the last *pre-evidence* position and training with *de-lexicalized* evidence both retain most of the measured gain, indicating that the transferred structure is present in the representation rather than only in the generated string. The probe protocol (splits, annotations, probe class), confidence intervals, and a shuffled-belief-target control are in Appendix C (Table 19).

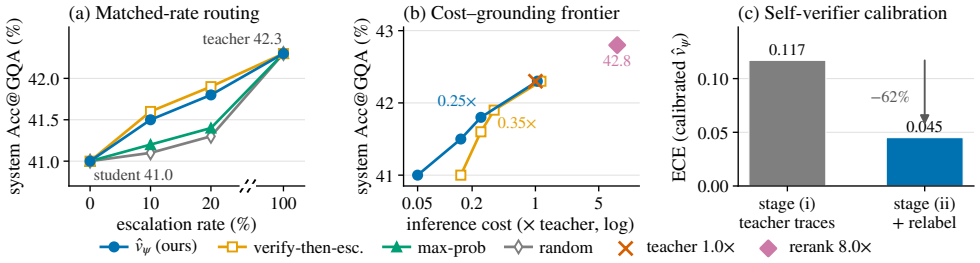

Figure 2: **Deployment analysis on NExT-GQA.** *(a)* $\hat{v}_\psi$-routing dominates max-probability and random routing and tracks head-free verify-then-escalate within 0.1. *(b)* Cost–grounding frontier: sweeping κ traces the cascade between student and teacher; verify-then-escalate pays the verifier on every input; reranking sits far off the frontier ($\kappa \to 1$ endpoint: 1.05×). *(c)* Stage-(ii) relabeling cuts the ECE of $\hat{v}_\psi$ from 0.117 to 0.045.

Calibration and cascade (Q5). After stage-(ii) relabeling $\hat{v}_\psi$ is well calibrated (ECE 0.045, from 0.117 with teacher-trace training alone; Figure 2c), and $\hat{v}_\psi$-routing dominates max-probability and random routing at every escalation rate (Table 5), matching verify-then-escalate within 0.1 Acc@GQA at 0.10× lower cost; the gain comes from where the head sends inputs (on the escalated 20% the teacher beats the student by 4.0 Acc@GQA; Table 16). The terms of Proposition 6 are tabulated there: with the student-side soundness term clipped at 1 the bound is vacuous through 50% escalation and loose at 80%, so the gains at the cost-driven operating points come from the routing signal's correlation with the student's *own* errors, which it does not model. Because $\hat{v}_\psi$ targets verifier acceptance, calibration to Y does not itself certify groundedness; Table 17 therefore reports $\hat{v}_\psi$'s discrimination of the joint correct-and-localized event and of correct-but-unlocalized outputs directly.

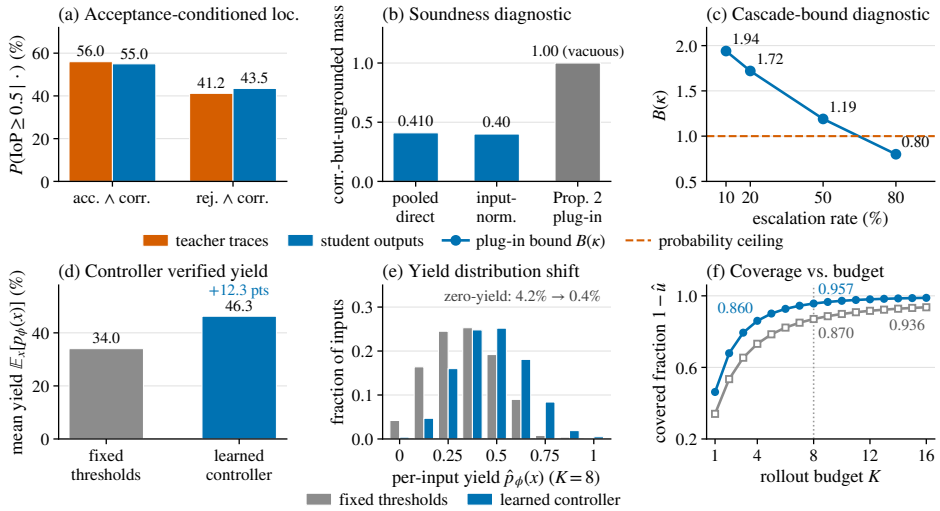

Figure 3: **Measuring the analysis (top) and the coverage channel (bottom) on NExT-GQA.** *(a)* Localization among correct traces vs. verifier acceptance: $41.2 \rightarrow 56.0$ (teacher), $43.5 \rightarrow 55.0$ (re-verified student; Table 14). *(b)* Direct correct-but-ungrounded mass (0.410 pooled, ≈ 0.40 input-normalized) vs. the Proposition 2 plug-in, which clips at 1 (Table 15). *(c)* Cascade bound $B(\kappa)$ vs. the measured gap of 0.40 (Table 16). *(d)* Mean verified yield at $K{=}8$, hand-set vs. learned controller (3 seeds). *(e)* Per-input yield at $K{=}8$: zero-yield mass $4.2\% \rightarrow 0.4\%$. *(f)* Coverage vs. K: the learned controller at $K{=}8$ (0.957) beats hand-set thresholds at $K{=}16$ (0.936).

Table 8: **Ablations** (means over three seeds; seed std $\leq$ 0.3 Acc@GQA, so differences below ≈ 0.5 are within noise). Grounding on NExT-GQA, accuracy on NExT-QA; defaults $\gamma{=}0.5$, $K{=}8$, answer-first $(y, \hat{g}, b)$, $\alpha{=}0$; GFS as in Table 4; controller variants in Table 9.

Configuration	NExT-QA Acc	Acc@GQA	GFS
Full VERA	81.2	41.0	0.66
fixed thresholds (no learned controller)	80.4	39.1	0.62
answer-only target ($\lambda_e{=}\lambda_{\hat{g}}{=}\lambda_b{=}\lambda_r{=}0$)	80.8	35.2	0.52
+evidence, no rationale/alignment ($\lambda_b{=}\lambda_r{=}0$)	81.0	40.3	0.64
no representation alignment ($\lambda_r{=}0$)	81.1	40.4	0.64
evidence-first serialization $(\hat{g}, y, b)$	81.0	40.8	0.66
no inverse-yield debiasing / same-sample propensities (no cross-fitting)	81.2 / 81.1	40.6 / 40.4	0.65 / 0.64
no tilt ($\gamma{=}0$) / aggressive tilt ($\gamma{=}1$)	80.6 / 80.9	39.8 / 40.1	0.63 / 0.64
rollout budget $K{=}4$ (vs. 8)	80.7	39.5	0.63
InfoNCE negatives: other-inputs only / rejected-trace only	81.2 / 81.1	40.7 / 40.6	0.65 / 0.64
verifier-only selection of $(\eta, \gamma, \kappa, \lambda_\cdot)$	81.2	40.8	0.65
coverage-sensitive controller / label-augmented reward ($\alpha{=}0.5$)	81.2 / 81.3	41.1 / 41.2	0.66 / 0.66

Ablations and generalization (Tables 8 and 7). In order of weight: typed-evidence targets (+5.1 Acc@GQA over answer-only; +5.8 for the full objective), the learned controller (+1.9 over fixed thresholds), rollout budget (+1.5 from $K{=}4$ to 8), and, within seed noise, representation alignment (+0.6; its main effect is in the probes of Table 6) and cross-fitted propensities (+0.6). Moderate tilt beats none or aggressive tilt; both negative sources in Eq. (7) help; verifier-only dial selection costs 0.2; evidence-first serialization costs 0.2, so answer-first is a convenience, not a load-bearing choice. Grounded accuracy rises with scale ($37.2/41.0/42.4$ at 1B/4B/12B; the 12B single-pass student reaches the single-greedy teacher's 42.3 at $0.09\times$ cost, with best-of-K reranking at $8.0\times$ the true upper envelope, 42.8) and the advantage over correctness filtering is $+3.8/+5.0/+5.5$ at the three scales (Table 7).

4 CONCLUSION

VERA's hypothesis is that typed, checked evidence supplies training information beyond answer correctness; the correctness-filtered controls, replicated on three benchmarks and two teacher–verifier pairs, test it. On NExT-GQA the correct-but-unlocalized mass is ≈ 0.40 for every system: the gains are in grounded accuracy and localization among correct answers, not in that mass, which our analysis bounds only vacuously. Limitations: the controller can exploit verifier failures; temporal overlap captures part of evidence quality; teacher-forced alignment can encode textual shortcuts; the head inherits verifier errors; trace collection pays off above $\approx 3{\times}10^5$ queries against the teacher (an order of magnitude more against verify-then-escalate; Appendix C).

Reproducibility Statement

Code for the controller, tilted filtering, distillation, self-verification head, cascade, and all evaluation scripts—including the exact NExT-GQA evaluator invocation, the GFS and question-swap scripts, and the per-question localization matrix behind §3—is provided in the attached ZIP archive and accompanying **Supplementary Material**. The LINGUA teacher is run with its authors' implementation; VideoMind is public. The supplementary package also includes the trace corpora, verifier labels, and the 1,000-clip instrumentation annotations used in our experiments. Hyperparameters and implementation details are provided in Appendix B and the Supplementary Material. All our-pipeline numbers on NExT-GQA are means over three seeds on the test split, touched once, while the replication cells of Table 2 are single runs (with paired-bootstrap CIs in Table 12), and re-evaluated baselines (‡) are single measured runs of released checkpoints.

Use of generative AI

Generative AI tools were used for editing and polishing the text only; all technical content, experiments, and analyses are the authors' own.

References

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2412.12075.

Jr-Jen Chen, Yu-Chien Liu, Yu-Hua Li, An-Chieh Chen, Chun-Yi Yeh, and Winston H. Hsu. ReXTime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. arXiv:2406.19392.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025b.

Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Weihong Lin, Zekun Wang, Bohan Zeng, Yang Shi, Sihan Yang, Qiang Liu, Pengfei Wan, et al. Vidbridge-r1: Bridging qa and captioning for rl-based video understanding models with intermediate proxy tasks. In *International Conference on Learning Representations*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jisheng Dang, Huilin Song, Junbin Xiao, Bimei Wang, Han Peng, Haoxuan Li, Xun Yang, Meng Wang, and Tat-Seng Chua. Mupa: Towards multi-path agentic reasoning for grounded video question answering. *arXiv preprint arXiv:2506.18071*, 2025.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4443–4458, 2020.

Xinxin Dong, Baoyun Peng, Haokai Ma, Yufei Wang, Zixuan Dong, Fei Hu, and Xiaodong Wang. Leadqa: Llm-driven context-aware temporal grounding for video question answering. In *International Conference on Intelligent Computing*, pp. 103–114. Springer, 2026.

Saman Forouzandeh, Wei Peng, Xinghuo Yu, and Mahdi Jalili. Bridging the grounding gap in VideoQA: Linguistic belief-state reasoning with typed memory. In *Forty-third International Conference on Machine Learning*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 14852–14882, 2023.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the association for computational linguistics: ACL 2023*, pp. 8003–8017, 2023.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. Eg-vqa: Benchmarking verifiable video question answering with grounded temporal evidence. *arXiv preprint arXiv:2606.24797*, 2026.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Yogesh Kumar. Catching hallucinated citations in video-llm question answering: A self-verification pipeline and verifier ablation study. *arXiv preprint arXiv:2608.15574*, 2026.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Tony Lee and Percy Liang. Self-verified distillation: Your language model is secretly its own synthetic data pipeline. *arXiv preprint arXiv:2605.26132*, 2026.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. IntentQA: Context-aware video intent reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11929–11940. IEEE, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Hao Lin, Kunyang Lv, Xu Jiang, Jingqi Tian, Zhongjing Du, Jiayu Ding, Qiaoman Zhang, and Hongbo Jin. Visd: Enhancing video reasoning via structured self-distillation. *arXiv preprint arXiv:2605.06094*, 2026.

Huabin Liu, Filip Ilievski, and Cees GM Snoek. Commonsense video question answering through video-grounded entailment tree reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3262–3271. IEEE, 2025.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, et al. Videoauto-r1: Video auto reasoning via thinking once, answering twice. *arXiv preprint arXiv:2601.05175*, 2026a.

Ye Liu, Kevin Qinghong Lin, Chang-Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, 2026b.

Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. VITED: Video temporal evidence distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2503.12855.

Romy Luo, Zihui Sherry Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. *Advances in Neural Information Processing Systems*, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13235–13245. IEEE, 2024.

Ahmad Mohammadshirazi, Pinaki Prasad Guha Neogi, Dheeraj Kulshrestha, and Rajiv Ramnath. Docval: Validated chain-of-thought distillation for grounded document vqa. *arXiv e-prints*, pp. arXiv–2511, 2025.

Tony Montes and Fernando Lozano. ViQAgent: Zero-shot video question answering via agent with open-vocabulary grounding validation. In *Proceedings of the 21st International Conference on Computer Vision Theory and Applications (VISAPP)*, 2026. arXiv:2505.15928.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Jinhwan Seo, Kyubeom Han, Jumin Lee, Junhyug Noh, and Sung-eui Yoon. What you ask is what you ground: Bridging question intent to temporal evidence for grounded videoqa. *arXiv preprint arXiv:2608.15708*, 2026.

Sumaira Shafiq, Naveed Ejaz, Munam Ali Shah, Rashid Kamal, Adnan Sohail, and Sheraz Aslam. Intra-video temporal-aware rag: A self-contained framework for video-based question answering. *Computers, Materials and Continua*, 88(2), 2026.

Sahil Shah, SP Sharan, Harsh Goel, Minkyu Choi, Mustafa Munir, Manvik Pasula, Radu Marculescu, and Sandeep Chinchali. Neus-qa: Grounding long-form video understanding in temporal logic and neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8805–8813, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Chendong Wang, Donglin Bai, Yifan Yang, Xiao Jin, Anlan Zhang, Rui Wang, Shiqi Jiang, Yuqing Yang, Hao Wu, Qi Dai, et al. Video-in-the-loop: Span-grounded long video qa with interleaved reasoning. *arXiv preprint arXiv:2510.04022*, 2025.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, 2026a.

Shijie Wang, Honglu Zhou, Ziyang Wang, Ran Xu, Caiming Xiong, Silvio Savarese, Chen Sun, and Juan Carlos Niebles. Evidence-backed video question answering. *arXiv preprint arXiv:2607.11862*, 2026b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575, 2023.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 2024.

Yuanxing Xu, Yuting Wei, Shuai Zhong, Xinming Chen, Jinsheng Qi, and Bin Wu. Exploring question guidance and answer calibration for visually grounded video question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3121–3133, 2024.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Mingfang Zhang, Jingjing Pan, Ashutosh Kumar, Rajat Saini, Mustafa Erdogan, Hsuan-Kung Yang, Caixin Kang, Yifei Huang, Yoichi Sato, and Quan Kong. Cast-bench: Benchmarking causal chain-grounded spatio-temporal reasoning for video question answering. *arXiv preprint arXiv:2605.23216*, 2026.

Shiding Zhu, Yudi Qi, Yajie Wang, Jiaze Li, Chao Song, Yaorui Shi, Yibo Miao, Hanqi Gao, and Kai Zhang. Escaping the self-confirmation trap: An execute-distill-verify paradigm for agentic experience learning. *arXiv preprint arXiv:2606.24428*, 2026.

Bo Zou, Chao Yang, Yu Qiao, Chengbin Quan, and Youjian Zhao. VideoDistill: Language-aware vision distillation for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2404.00973.

A Theoretical Analysis and Proofs

We formalize *why* verification—not the chain-of-thought format—makes the distilled student grounded, taking care to avoid circularity: every quantity in our bounds is measurable from the teacher's logs, estimable on a small annotated instrumentation subset, or a standard learning-theoretic term—and we flag explicitly which assumptions are verifier-side, which are population-level, and which are estimated rather than guaranteed. We also state, at each step, which results hold for a *generic* verifier-equipped teacher (i.e., are not specific to LINGUA) and which are specific to our filtering-and-distillation construction, since this distinction is what lets Corollary 1 transfer the bounds to other teachers. Statements and discussion come first (§A.1–A.9); full proofs are collected in §A.11 at the end of this section.

Summary of results. Proposition 2 bounds the correct-but-ungrounded mass of the rejection-filtered teacher by a quantity δ built from the verifier's per-step false-accept rate ε_v and boundary-imprecision mass $\varepsilon_{\mathrm{loc}}$ against the teacher's raw grounded fraction and verified yield; tilting inflates it by at most $e^{\gamma(1-\eta)}$, and a student within TV-distance ε of the tilted law inherits $G \le e^{\gamma(1-\eta)}\delta + \varepsilon$ (Lemma 1, Proposition 3). Theorem 1 shows no such bound exists for raw-trace distillation, Proposition 4 makes coverage explicit at the input-cluster level, and Proposition 6 bounds the cascade without a fidelity term. All are population-level diagnostics; on NExT-GQA the soundness plug-in saturates at 1, and the cascade plug-in is vacuous through 50% escalation and loose at 80% (Remark 4).

A.1 Setup

Inputs $x = (V, q) \sim \mathcal{D}$. A model π induces an output law over $(\hat{y}, \hat{g})$; evidence *localizes* the gold region if $\mathrm{loc}(\hat{g}, g^\star) = 1$, where loc is the fixed criterion of §2. Every statement below holds for *either* instantiation—$IoP \ge 0.5$ (the official Acc@GQA rule) or the stricter $IoU \ge 0.5$—provided the

same criterion is used to define the primitives of Assumption 1 and to report the metric; the measured constants of course differ between them, and we report both. Define $\mathrm{Correct}(\pi) = \mathbb{E}_x\mathbb{P}_\pi[\hat{y} = y^\star]$, $\mathrm{Grounded}(\pi) = \mathbb{E}_x\mathbb{P}_\pi[\hat{y} = y^\star \wedge \hat{g}\ \text{localizes}\ g^\star]$, and the *grounding gap* $G(\pi) = \mathrm{Correct}(\pi) - \mathrm{Grounded}(\pi) \geq 0$. Let $\pi_{\mathrm{T}}^{\mathrm{Ver}} = \mathcal{R}_{\eta,\gamma}(\pi_{\mathrm{T}})$ with induced output law $P^{\mathrm{Ver}}(\cdot \mid x)$.

A.2 Unification of prior filtering strategies

Before analyzing $\mathcal{R}_{\eta,\gamma}$'s soundness, we make precise the sense in which it generalizes the filtering strategies used by the closely related distillation recipes discussed in §1, so that the novelty of VERA is stated as a property of the operator family rather than asserted narratively.

Let a *filtering strategy* be a pair (ϕ, ω) of a retention predicate $\phi : \mathcal{T} \to \{0, 1\}$ and a weight function $\omega : \mathcal{T} \to \mathbb{R}_{\geq 0}$ applied to the teacher's raw trace distribution, $\pi(\tau \mid x)\phi(\tau)\omega(\tau)$ (renormalized). Recall $S(\tau) = \frac{1}{T}\sum_t v_t \in [0, 1]$ is our *step-level* verified fraction over typed postcondition/temporal checks (Eq. (1)), as opposed to a trace-level binary correctness or verifier-acceptance label.

Proposition 1 (Unification of filtering strategies)**.** *The tilted filtering operator $\mathcal{R}_{\eta,\gamma}$ of Eq.* (4) *recovers, as special cases:*

1. ***Unfiltered chain-of-thought distillation*** *(the base case of System-2 distillation, (Yu et al., 2024)): $(\phi, \omega) \equiv (1, 1)$, obtained from $\mathcal{R}_{\eta,\gamma}$ at $\eta = 0, \gamma = 0$, since $S(\tau) \geq 0$ holds identically and $e^{0 \cdot S(\tau)} = 1$.*
2. ***Correctness-only self-training*** *(STaR, (Zelikman et al., 2022)): $\phi(\tau) = \mathbf{1}[\hat{y} = y^\star]$, $\omega \equiv 1$. This is $\mathcal{R}_{\eta=1,\gamma=0}$ applied with the score function $S(\tau)$ redefined as $S_{\mathrm{corr}}(\tau) \coloneqq \mathbf{1}[\hat{y} = y^\star] \in \{0, 1\}$ in place of the step-averaged verification fraction; i.e., STaR occupies the hard-threshold boundary of the family under a coarser (trace-level, verifier-free) score.*
3. ***Hard verifier-gated retention*** *(V-STaR-style filtering, (Hosseini et al., 2024); single-round instances of execute–distill–verify, (Zhu et al., 2026); self-verified synthetic-data curation, (Lee & Liang, 2026)): $\phi(\tau) = \mathrm{Ver}(\tau)$, $\omega \equiv 1$. This is exactly $\mathcal{R}_{\eta,\gamma=0}$: hard rejection at a verifier threshold with no tilt.*

Consequently, at $\gamma = 0$ the family $\{\mathcal{R}_{\eta,0}\}_{\eta\in\{0,1\}}$ already spans the two boundary points—no filtering and hard verifier-gated filtering—occupied by cases 1 and 3 above, and case 2 is the same boundary point under a different (verifier-free) scoring function. VERA*'s filtering choice differs along two axes not spanned by any of the three: (a) an independent, continuously tunable tilt $\gamma > 0$ that reweights* within *the certified region rather than only thresholding it, and (b) a score $S(\tau)$ defined at the level of individually typed, checkable per-step postconditions rather than a single trace-level or answer-level label, which is what subsequently permits offline re-verification of* student *outputs (§2.4) and the step-level soundness argument of §A.3.*

The proof (§A.11) is definitional substitution into Eq. (4) and is given there for completeness. We stress what Proposition 1 does and does not claim: it does not claim that (Hosseini et al., 2024), (Zhu et al., 2026), or (Lee & Liang, 2026) literally implement $\mathcal{R}_{\eta,0}$ with our particular step-typed $S(\tau)$—their verifiers, training loops, and target domains differ in ways this abstraction elides. It claims only that, restricted to the filtering-operator level, their retention rules are representable as boundary instances of the same two-parameter family. Two further limits on how much weight this can bear. Case 2 is obtained by *substituting a different score function* into the operator, not by choosing parameters within it, so calling STaR a special case is a statement about the shape of a retention rule and nothing more. And a retention rule is a small part of a training algorithm: none of these methods is reconstructed by our operator, because their sampling, optimization, and target domains differ. We accordingly treat Proposition 1 as organizing the design space, not as a novelty claim, and rest the novelty claim instead on the combination named in §1, evaluated in §3. §3 uses this characterization directly: our "Ver only" vs. "correctness only" vs. "Ver$\wedge$correct" ablation of §2.2 instantiates cases 2 and 3 above as matched-mechanism baselines inside our own pipeline, rather than relying solely on each method's originally reported numbers under different backbones and compute budgets.

A.3 Soundness from measurable verifier primitives

Rather than *assuming* that verified-and-correct traces are grounded (which would come close to assuming our conclusion), we express the grounding gap of the filtered teacher through perception-

level properties estimable from the teacher's instrumentation (V1) and, for V2, on an annotated instrumentation subset used only for measurement. We state the primitives as *input-dependent* conditional probabilities, which always exist, rather than as uniform constants: the bound then holds with no uniformity requirement across inputs, and the constant-primitive form is the special case in which the per-input quantities are replaced by their suprema. What is assumed is therefore not that the primitives are small, but only the two conditional-probability *structures* below; what is estimated is their averaged magnitude. The result should be read as a conditional diagnostic bound under population-level, measured assumptions, not as an unconditional guarantee.

Assumption 1 (Verifier primitives, input-dependent form)**.** Fix a margin $\rho < \eta$. For each input x in the support of $\mathcal{D}$, write $U_{\mathrm{lo}}(x)$ for the class of teacher traces on x that are correct, unlocalized, and have true-support fraction at most ρ (defined in the proof of Proposition 2), and define the two *per-input primitives*

$$\varepsilon_v(x) \;\coloneqq\; \sup_t \; \mathbb{P}\big(v_t = 1 \,\big|\, \text{step } t \text{ unsupported},\; x,\; \tau \in U_{\mathrm{lo}}(x)\big),$$

$$\varepsilon_{\mathrm{loc}}(x) \;\coloneqq\; \mathbb{P}_{\pi_{\mathrm{T}}}\big[\text{support}(\tau) > \rho \,\big|\, \hat{y} = y^\star,\; \neg\text{localized},\; x\big].$$

(V1, conditional step soundness) $\varepsilon_v(x)$ bounds the false-accept probability of an *unsupported* step conditionally on the class $U_{\mathrm{lo}}(x)$. We state the conditional form deliberately: a marginal per-step false-accept rate does *not* imply the conditional one, because detector failures may correlate with video difficulty, action type, or trace length. Accordingly, $\varepsilon_v(x)$ is a function of the event detector's false-positive rate and the semantic-similarity threshold on x, and we estimate it on annotated clips *stratified* by action type, trace-length tercile, and question difficulty; only linearity of expectation is used—no independence across steps. The resulting plug-in bound saturates, so Table 15 reports the aggregate diagnostic. *(V2, boundary-imprecision mass)* $\varepsilon_{\mathrm{loc}}(x)$ is the conditional probability, among correct-but-ungrounded traces on x, of the traces that are semantically supported by real events but cite temporally imprecise spans. Both quantities are conditional probabilities and therefore always exist in $[0, 1]$; the assumption is only that the events they refer to are well defined (a true-support fraction and a localization criterion are fixed, §A.3). The *constant-primitive form* is the special case $\varepsilon_v \coloneqq \sup_x \varepsilon_v(x)$, $\varepsilon_{\mathrm{loc}} \coloneqq \sup_x \varepsilon_{\mathrm{loc}}(x)$; nothing below requires it.

V2 deserves a comment: it is exactly the boundary-imprecision failure mode our agentic teacher's instrumentation already quantifies—answers whose postconditions verify against real events while the cited boundaries miss the IoU threshold—so $\varepsilon_{\mathrm{loc}}(x)$ is not hypothetical; it is a population statistic the teacher's instrumentation already estimates. All *other* ungrounded traces have support at most ρ by definition of the split, which is what avoids the unsatisfiable requirement that *all* ungrounded traces be support-deficient. One measurability distinction matters for interpreting "label-free": $\varepsilon_v(x)$ is a verifier-side quantity, but estimating $\varepsilon_{\mathrm{loc}}(x)$, $g_{\mathrm{T}}(x)$, and the direct gap $\hat{\delta}$ requires gold spans on an annotated instrumentation subset. These annotations enter only the *measurement* of the bound; the training pipeline—controller reward, filtering, distillation, and self-verifier supervision—consumes no gold labels. Three further honesty notes. First, V2 is a *population-level* statement about the teacher's trace distribution, not a verifier-side statistic: estimating $\varepsilon_{\mathrm{loc}}(x)$ requires gold temporal evidence together with an operational definition of the true-support fraction, which we fix as the fraction of steps whose declared postconditions are satisfied by *gold-aligned* events (detected events matched to gold spans at IoU $\geq \eta_g$; Appendix C). Second, $\varepsilon_{\mathrm{loc}}$ can dominate δ even when the postcondition checker is nearly exact—a low false-accept rate does not by itself certify temporal localization—so we report the measured decomposition of δ and acceptance-conditioned localization statistics (Table 14) rather than the bound alone, and we describe the bound as expressed in *measured* primitives rather than as purely label-free. Third, per-input primitives are estimated from K rollouts per input, so individual estimates are coarse; the quantity Proposition 2 actually needs is their *average* over inputs of a clipped ratio. That plug-in reaches the clip at the probability ceiling; Table 15 therefore reports the aggregate direct estimate and the fact that the plug-in is vacuous.

Proposition 2 (Soundness bound from measured verifier primitives)**.** *Let* $g_{\mathrm{T}}(x) = \mathbb{P}_{\pi_{\mathrm{T}}}[\hat{g} \text{ localizes } g^\star \mid \hat{y} = y^\star, x]$ *be the raw teacher's grounded fraction among correct answers on input* x *and* $\bar{q}(x) = \mathbb{P}_{\pi_{\mathrm{T}}}[\mathrm{Ver}(\tau) = 1 \mid \hat{y} = y^\star, x]$ *its verified yield among correct answers on* x*. Define the per-input bound*

$$\delta(x) \;\coloneqq\; \min\left\{1,\; \frac{1 - g_{\mathrm{T}}(x)}{\bar{q}(x)}\left(\varepsilon_{\mathrm{loc}}(x) + \frac{\varepsilon_v(x)}{\eta - \rho}\right)\right\}$$

on inputs with $\mathbb{P}_{\pi_{\mathrm{T}}}[\hat{y} = y^\star \wedge \mathrm{Ver} = 1 \mid x] > 0$*, and* $\delta(x) \coloneqq 0$ *otherwise (such inputs contribute nothing to* $G(\pi_{\mathrm{T}}^{\mathrm{Ver}})$*: if no correct trace is certified, the filtered law on* x *carries no correct-and-ungrounded mass, and if no trace at all is certified,* x *is uncovered and is accounted for by* u *in Proposition 4). Under Assumption 1, the rejection-filtered teacher (*$\gamma = 0$*) satisfies*

$$G(\pi_{\mathrm{T}}^{\mathrm{Ver}}) \;\le\; \delta \;\coloneqq\; \mathbb{E}_x\big[\delta(x)\big].$$

In the constant-primitive special case $\varepsilon_v \coloneqq \sup_x \varepsilon_v(x)$*,* $\varepsilon_{\mathrm{loc}} \coloneqq \sup_x \varepsilon_{\mathrm{loc}}(x)$*, write* $c \coloneqq \varepsilon_{\mathrm{loc}} + \varepsilon_v/(\eta-\rho)$ *and* $\bar{r}_{\mathrm{u}} \coloneqq \mathbb{E}_x\big[(1-g_{\mathrm{T}}(x))/\bar{q}(x)\big]$ *for the* unclipped *mean ratio; then, since* $t \mapsto \min\{1,t\}$ *is concave, Jensen gives*

$$\delta \;=\; \mathbb{E}_x\big[\,\min\{1,\, \tfrac{1-g_{\mathrm{T}}(x)}{\bar{q}(x)}c\}\big] \;\le\; \min\big\{1,\; \bar{r}_{\mathrm{u}}\, c\big\}.$$

The clip must stay around the whole *product: the tempting form* $\mathbb{E}_x[\min\{1,(1-g_{\mathrm{T}}(x))/\bar{q}(x)\}]\cdot c$ *is* not *an upper bound, because* $\min\{1,ac\} \le \min\{1,a\}c$ *is false in general (*$a = 2$*,* $c = 0.1$ *gives* 0.2 *on the left and* 0.1 *on the right). When* g_{T} *and* $\bar{q}$ *are themselves constant the bound reduces to the global form* $\min\{1, c\,(1-g_{\mathrm{T}})/\bar{q}\}$*. The clip at* 1 *is not cosmetic: on inputs with small verified yield among correct traces the unclipped ratio can exceed* 1*, but a conditional probability cannot, so the clipped form is both tighter and finite on every input.*

Remark 1 (Estimation of δ). δ is the mean over inputs of a clipped per-input ratio, and is estimated by plugging in per-input estimates of $(g_{\mathrm{T}}(x), \bar{q}(x), \varepsilon_{\mathrm{loc}}(x), \varepsilon_v(x))$ from the K rollouts and annotated evidence of each instrumentation input, then averaging; bootstrap confidence intervals resample inputs. Two consequences of the clip matter for reading the estimate. First, because $\delta(x) \le 1$ pointwise, the plug-in estimate of δ is finite and its bootstrap variance is bounded, whatever the estimated $\bar{q}(x)$; the same is not true of the unclipped $\bar{r}$, which can be dominated by a few low-yield inputs. Second, the *direct* estimate $\hat{\delta}$ that we compare the bound against must weight inputs the same way $G(\pi_{\mathrm{T}}^{\mathrm{Ver}})$ does—equally, with the per-input filtered law normalized within the input—rather than pooling accepted-correct traces across inputs (which weights inputs by their accepted-correct mass). Table 14 reports the pooled statistic for interpretability and Table 15 the input-normalized one that the proposition bounds, together with the fraction of inputs on which the clip binds and the distribution of the per-input terms.

Corollary 1 (Teacher-agnostic soundness)**.** *Proposition 2 is stated entirely in terms of a generic trace distribution* $\pi_{\mathrm{T}}(\tau \mid x)$ *satisfying Assumption 1 and the interface of §2 (typed postconditions, a per-step verification indicator* v_t*, and the aggregate score* $S(\tau)$*); no step of its proof (§A.11) uses a property specific to LINGUA. Consequently, for* any *verifier-equipped agentic teacher* π'_{T} *instantiating the same interface, with its own per-input primitives* $(\varepsilon'_v(x), \varepsilon'_{\mathrm{loc}}(x))$ *and margin* ρ' *under Assumption 1, Propositions 2, 3, and 4 and Theorem 1 hold verbatim with* $\pi_{\mathrm{T}}, \varepsilon_v(\cdot), \varepsilon_{\mathrm{loc}}(\cdot), \rho$ *replaced by* $\pi'_{\mathrm{T}}, \varepsilon'_v(\cdot), \varepsilon'_{\mathrm{loc}}(\cdot), \rho'$*. LINGUA is the instantiation from which we* measure *these constants for our experiments, not a hypothesis of the theorems. Being generic in the teacher is a property of the* bound*, not evidence that the empirical effect transfers; that is tested separately by rebuilding the pipeline on the public agentic teacher VideoMind (Liu et al., 2026b) and re-measuring the primitives for it (Table 2).*

Corollary 2 (Sensitivity to verifier quality)**.** *Fix* η, ρ *and the teacher's raw rates* $g_{\mathrm{T}}(\cdot), \bar{q}(\cdot)$*. Let* δ *and* δ' *be the bounds of Proposition 2 under two verifiers with primitives* $(\varepsilon_v(\cdot), \varepsilon_{\mathrm{loc}}(\cdot))$ *and* $(\varepsilon'_v(\cdot), \varepsilon'_{\mathrm{loc}}(\cdot))$*. Since* $t \mapsto \min\{1,t\}$ *is* 1*-Lipschitz and the unclipped per-input ratio is affine in the primitives,*

$$|\delta-\delta'| \;\le\; \mathbb{E}_x\left[\frac{1-g_{\mathrm{T}}(x)}{\bar{q}(x)}\Big(|\varepsilon_{\mathrm{loc}}(x)-\varepsilon'_{\mathrm{loc}}(x)| + \tfrac{1}{\eta-\rho}|\varepsilon_v(x)-\varepsilon'_v(x)|\Big)\right] \;\le\; \bar{r}_{\mathrm{u}}\Big(\|\varepsilon_{\mathrm{loc}}-\varepsilon'_{\mathrm{loc}}\|_\infty + \tfrac{1}{\eta-\rho}\|\varepsilon_v-\varepsilon'_v\|_\infty\Big),$$

with $\bar{r}_{\mathrm{u}} \coloneqq \mathbb{E}_x[(1-g_{\mathrm{T}}(x))/\bar{q}(x)]$ *the unclipped ratio (the middle expression is the sharper, input-weighted form). The corollary provides a sensitivity characterization of the soundness bound with respect to verifier quality: changes in the false-accept and boundary-imprecision primitives are propagated through the input-weighted expression above. Appendix D evaluates this relationship empirically by perturbing the real verifier's perception pipeline and reporting the measured grounding response under those perturbations.*

Lemma 1 (Tilt robustness)**.** *Tilting is* not *innocuous:* $e^{\gamma S}$ *up-weights high-*S *traces, which among ungrounded survivors are precisely the false-accept-heavy ones, so soundness need not improve monotonically in* γ*. It is, however, controlled: since* $\bar{S}(\tau) \in [\eta, 1]$ *on the surviving set, tilt weights there differ by at most* $e^{\gamma(1-\eta)}$*, and hence*

$$G\big(\mathcal{R}_{\eta,\gamma}(\pi_{\mathrm{T}})\big) \;\le\; \delta_\gamma \;\coloneqq\; e^{\gamma(1-\eta)}\,\delta.$$

Every downstream result that targets the tilted law is stated in terms of δ_γ, not δ; at $\gamma = 0$ the two coincide, and at our operating points ($\eta = 0.7$, $\gamma \in \{0, 0.5, 1\}$) the inflation factor is 1, 1.16*, and* 1.35.

Remark 2 (What the ceiling does and does not imply). Lemma 1 is an *upper* bound that grows with γ. A growing ceiling does not imply that the measured gap must grow: it only ceases to forbid it. Accordingly we report the measured gap of the tilted teacher against the ceiling as a consistency check (does the measurement stay below the bound?), and treat both the observed benefit of moderate tilt and the observed degradation at aggressive tilt as empirical findings the ceiling permits but does not entail.

The proof of Proposition 2 (§A.11) is a Markov bound—an ungrounded, support-deficient trace passing the η filter must accumulate at least an $(\eta - \rho)$-fraction of false accepts—followed by Bayes' rule, with the boundary-imprecise mass handled separately through V2. The bound exposes exactly *which* verifier properties matter: soundness degrades linearly in the per-step false-accept rate and in the boundary-imprecision mass (until the clip binds), is amplified when the filter margin $\eta - \rho$ is thin, and improves with the teacher's raw grounding and verified yield. Every term bottoms out in quantities LINGUA reports or that the instrumentation subset estimates, so δ is an *estimated* quantity with a confidence interval rather than a leap of faith; we report its empirical decomposition, its per-input distribution, and the bound-vs-direct-estimate comparison in §3 and Appendix C.

A.4 Distillation transfers the bound; raw traces provably cannot

Assumption 2 (Distillation fidelity)**.** $\mathbb{E}_x\big[\mathrm{TV}(\pi_\theta(\cdot \mid x), P^{\mathrm{Ver}}(\cdot \mid x))\big] \le \varepsilon$.

Proposition 3 (Grounding-gap bound)**.** *Under Assumptions 1 and 2, $G(\pi_\theta) \le \delta_\gamma + \varepsilon$, with $\delta_\gamma = e^{\gamma(1-\eta)}\delta$ from Lemma 1 and δ as in Proposition 2. Two points deserve emphasis. The target of Assumption 2 is $P^{\mathrm{Ver}} = \mathcal{R}_{\eta,\gamma}(\pi_{\mathrm{T}})$, the* tilted *law, so the soundness term must carry the tilt factor; using δ here would bound the wrong distribution whenever $\gamma > 0$. And the transfer term is ε, not 2ε, by the sharpened form of Lemma 3.*

Prior work on filtered distillation leaves the contrast with raw-CoT distillation informal. We make it a theorem.

Theorem 1 (Separation)**.** *For every $m \in (0, 1)$ there exists an input distribution, a teacher π_{T} with $G(\pi_{\mathrm{T}}) = m$, and a verifier satisfying Assumption 1 with $\delta = 0$, such that (i) the student distilled from $\mathcal{R}_{\eta,\gamma}(\pi_{\mathrm{T}})$ with fidelity ε has $G \le \varepsilon$, while (ii)* every *student matching the raw teacher with fidelity ε has $G \ge m - \varepsilon$. Consequently, no bound of the form $G(\pi_\theta) \le f(\delta, \varepsilon)$ with f independent of $G(\pi_{\mathrm{T}})$ can hold for raw-trace distillation.*

The construction (§A.11) is a two-trace teacher: on each input it emits, with probability m, a correct answer citing an irrelevant span (never verified) and otherwise a correct grounded trace (always verified). Perfect imitation of the raw mixture inherits the gap m exactly; filtering deletes it. The theorem upgrades "verification is the active ingredient" from a remark to a formal impossibility statement about the alternative.

A.5 Finite-sample coverage: what the controller buys, quantitatively

Population fidelity ε hides the mechanism by which the controller helps. Let N inputs be sampled i.i.d. from $\mathcal{D}$, each given budget K; call an input *covered* if at least one of its K traces is verified, let $\mathcal{C}$ be the set of covered inputs, $N_{\mathrm{cov}} = |\mathcal{C}|$, and let $u = \mathbb{E}_x\big[(1 - p_\phi(x))^K\big]$ be the uncovered mass. Outputs $o = (\hat{y}, \hat{g}, b)$ live in a finite space (multiple-choice answers $\times$ discretized span bins $\times$ a bounded rationale vocabulary), the student class $\mathcal{F}$ has log-loss $\ell_\theta(x, o) = -\log \pi_\theta(o \mid x) \in [0, B]$ and finite cardinality (or metric entropy, absorbed into constants).

The unit of analysis is the input, not the trace. Matching the construction of Eq. (5), a covered input i retains its $m_i \ge 1$ accepted traces $\{o_{ij}\}_{j \le m_i}$ with within-input tilt weights $\omega_{ij} \propto e^{\gamma S(\tau_{ij})}$, $\sum_j \omega_{ij} = 1$, and contributes the *cluster loss* $\bar{\ell}_\theta(x_i) \coloneqq \sum_j \omega_{ij}\, \ell_\theta(x_i, o_{ij}) \in [0, B]$. The student is the minimizer of the weighted empirical risk $\hat{L}(\theta) = \frac{1}{N_{\mathrm{cov}}} \sum_{i \in \mathcal{C}} \hat{w}_i\, \bar{\ell}_\theta(x_i)$. The traces within a cluster are dependent (they share x_i, are selected jointly, and are conditioned on $m_i \ge 1$); the analysis never treats them as independent samples—concentration is over the independent clusters, and the within-cluster average enters only through its conditional mean.

Weights, precisely. The ideal (Horvitz–Thompson) weight of a covered input and its estimated counterpart are

$$w^{\mathrm{HT}}(x) \coloneqq \frac{1/h(p_\phi(x))}{\mathbb{E}_{x'|C}\big[1/h(p_\phi(x'))\big]}, \qquad \hat{w}(x) \coloneqq \frac{\min\{W_{\max},\, 1/h(\hat{p}_\phi(x))\}}{\hat{Z}},$$

both normalized to have mean 1 over covered inputs ($\hat{p}_\phi$ is the cross-fitted propensity of §2.2). One subtlety must be handled rather than glossed: if $\hat{Z}$ is the *empirical* mean of the clipped weights over the covered sample, then the summands $\hat{w}_i \bar{\ell}_\theta(x_i)$ share a common random factor $1/\hat{Z}$ and are no longer independent, so cluster-level concentration does not apply to them directly. We therefore define $\hat{w}$ with the *deterministic* normalizer $Z \coloneqq \mathbb{E}_{x|C}[\min\{W_{\max}, 1/h(\hat{p}_\phi(x))\}]$, estimated out-of-fold, which restores independence across clusters; the deviation between the two is charged explicitly as

$$\varepsilon_Z \;\coloneqq\; \mathbb{E}\big|\hat{Z}/Z - 1\big| \;\cdot\; B\, W_{\max}/Z,$$

measured on the same folds and reported alongside the effective sample size. An alternative that avoids ε_Z entirely is to minimize the unnormalized weighted risk and rescale afterwards; we use the normalized form because it is better conditioned in practice, and pay the term. Under w^{HT} the weighted risk $L(\theta) \coloneqq \mathbb{E}_{x|C}\big[w^{\mathrm{HT}}(x)\, \mathbb{E}_{o\sim P^{\mathrm{Ver}}(\cdot|x)} \ell_\theta(x,o)\big]$ equals the P^{Ver} cross-entropy under the coverage-debiased input law (Remark 6). Two bias terms are then defined and measured:

$$\varepsilon_w \;\coloneqq\; \mathbb{E}_{x|C}\big|\hat{w}(x) - w^{\mathrm{HT}}(x)\big| \qquad \text{and} \qquad \varepsilon_{\mathrm{sn}} \;\coloneqq\; \sup_{\theta\in\mathcal{F}} \mathbb{E}_{x|C}\Big[w^{\mathrm{HT}}(x)\,\Big|\mathbb{E}\big[\bar{\ell}_\theta(x) \,\big|\, x,\, C(x)\big] - \mathbb{E}_{o\sim P^{\mathrm{Ver}}(\cdot|x)}\ell_\theta(x,o)\Big|\Big],$$

the propensity-estimation-plus-clipping error and the *self-normalization bias* of the within-cluster average, respectively. Since $\hat{w}$ and w^{HT} are both mean-one, ε_w is a genuine discrepancy between two normalized weight functions rather than a scale artifact. At $\gamma = 0$ the accepted traces of an input are, conditionally on the accepted index set, i.i.d. draws from $\mathcal{R}_{\eta,0}(\pi_{\mathrm{T}})(\cdot \mid x) = P^{\mathrm{Ver}}(\cdot \mid x)$, so the within-cluster average is conditionally unbiased and $\varepsilon_{\mathrm{sn}} = 0$ exactly; for $\gamma > 0$ the self-normalized tilt introduces the usual $O(1/m_i)$ bias, bounded by $e^{2\gamma(1-\eta)} B\, \mathbb{E}[1/m_i]$ (Remark 6) and estimated directly by comparing K-rollout cluster averages against high-budget rollouts on a probe subset.

Proposition 4 (Finite-sample grounding bound, cluster level)**.** *Let $\hat{\theta}$ minimize $\hat{L}$ over $\mathcal{F}$ and let $\varepsilon_{\mathrm{apx}} \coloneqq \min_{\theta\in\mathcal{F}} \mathbb{E}_{x|C}\big[w^{\mathrm{HT}}(x)\, \mathrm{KL}(P^{\mathrm{Ver}}(\cdot \mid x) \,\|\, \pi_\theta(\cdot \mid x))\big]$ be the class's weighted KL-approximation error to P^{Ver}. Conditionally on N_{cov} and on the cross-fitting folds, with probability at least $1-\zeta$ over the draw of the covered clusters,*

$$G(\pi_\theta[\hat{\theta}]) \;\le\; \delta_\gamma + u + \sqrt{\tfrac{\varepsilon_{\mathrm{apx}}}{2}} + \sqrt{B\,\varepsilon_w} + \sqrt{\varepsilon_{\mathrm{sn}}} + \sqrt{\varepsilon_Z} + C\,\big(W_{\max} B\big)^{1/2} \Big(\tfrac{\log|\mathcal{F}| + \log(F/\zeta)}{N_{\mathrm{cov}}}\Big)^{1/4},$$

where F is the number of cross-fitting folds and C is universal. Moreover $\mathbb{E}[N_{\mathrm{cov}}] = N\,\mathbb{E}_x\big[1 - (1-p_\phi(x))^K\big]$ and u are, respectively, strictly increasing and strictly decreasing in each $p_\phi(x)$—a property of these functionals, not *a claim about what the controller's objective optimizes (see the discussion below)—and the label-free controller objective satisfies $\mathbb{E}_x\mathbb{E}_{\tau\sim\pi_\phi}[\mathrm{Ver}(\tau)] = \mathbb{E}_x[p_\phi(x)]$. Neither u nor N_{cov} depends on γ: coverage is determined by $\{\mathbf{1}[\bar{S}(\tau) \ge \eta]\}$ alone. The soundness term carries the tilt factor of Lemma 1, and $\varepsilon_{\mathrm{apx}}$ is itself γ-dependent, since the target law P^{Ver} moves with γ.*

Three remarks keep the claim honest. First, the estimation rate for *total variation* is $N_{\mathrm{cov}}^{-1/4}$, not $N_{\mathrm{cov}}^{-1/2}$: per-input output laws are not observable from single clusters, so the rigorous route is weighted MLE $\to$ KL generalization at rate $N_{\mathrm{cov}}^{-1/2} \to$ Pinsker, which halves the exponent; the familiar $N^{-1/2}$ holds at the KL level and in the realizable case. Second, the concentration step is over *clusters*: the number of retained traces per input and their dependence structure affect the bound only through the conditional mean of the cluster loss, whose bias is $\varepsilon_{\mathrm{sn}}$; the multi-trace construction therefore costs nothing at $\gamma = 0$ and an explicit, measured term for $\gamma > 0$, but it does not buy a faster rate either—additional traces from the same input sharpen the cluster mean, not the count of independent clusters. Third, coverage enters the bound *twice*—as the explicit uncovered mass u and through N_{cov} in the estimation term. Both improve when p_ϕ rises *pointwise*, but the label-free objective of Eq. (2) maximizes only the *average* verified utility: an objective-improving update can in principle raise mean yield while lowering yield on difficult inputs, thereby *increasing* u on exactly the inputs for which filtering is most problematic. We therefore do not state pointwise improvement as a consequence of the objective. A sufficient condition is *no-degradation*—the update decreases no

$p_\phi(x)$—under which any strict increase in mean yield strictly decreases u; empirically, Figure 3(e) shows the learned controller shifting mass away from zero-yield inputs on our evaluation distribution, which is an *observed* property, not a theorem, and Appendix B describes a coverage-sensitive variant that optimizes $-\mathbb{E}_x[(1-\hat{p}_\phi(x))^K]$ directly (matched within noise; Table 8). This links Contribution 1 to Contribution 2 through a sample-complexity channel that can be checked empirically by plotting distilled grounded accuracy against measured (u, N_{cov}) under controller ablations.

A.6 Label-free selection of the filtering dials

The two dials enter Proposition 4 in different ways and must be kept apart. For fixed rollouts, the coverage event $C(x) = \{\exists j : S(\tau^{(j)}) \geq \eta\}$ depends on η *only*: the tilt γ reweights traces inside the certified region and cannot change whether an input has at least one certified trace. Hence u, N_{cov}, and the $N_{\mathrm{cov}}^{-1/4}$ estimation term are functions of η alone, computable from $\{\mathbf{1}[S(\tau) \geq \eta]\}$ with no gold spans or answers, and a label-free rule for η can target them exactly. The bound's dependence on γ enters in four places. Three of them provably *worsen* as γ grows: the tilt ceiling $\delta_\gamma = e^{\gamma(1-\eta)}\delta$ (Lemma 1), the self-normalization bias $\varepsilon_{\mathrm{sn}}$, and the effective sample size of the within-cluster weights. The fourth is the approximation error $\varepsilon_{\mathrm{apx}}$, which is *not* sign-determined: tilting moves the target law P^{Ver}, and a class $\mathcal{F}$ may approximate the tilted law better or worse than the untilted one. We therefore do not claim that the bound is monotone in γ, only that its *explicit* γ-dependent terms are non-decreasing, so nothing in the analysis recommends $\gamma > 0$. The empirical benefit of moderate tilt (Table 8) is a finding the analysis does not explain, and the label-free rule we use for γ is a heuristic that we state as such.

Definition 1 (Label-free dial selection)**.** Let $\hat{u}(\eta) \coloneqq \frac{1}{N}\sum_i \mathbf{1}\big[\max_{j\leq K} S(\tau_i^{(j)}) < \eta\big]$ be the empirical uncovered fraction at floor η, and, for a fixed η, let $\mathrm{ESS}(\gamma) \coloneqq \big(\sum_{i\in\mathcal{C}} \hat{w}_i\big)^2 / \sum_{i\in\mathcal{C}} \hat{w}_i^2 \cdot \mathrm{ESS}_{\mathrm{in}}(\gamma)$ be the effective sample size of the weighted corpus, where $\mathrm{ESS}_{\mathrm{in}}(\gamma)$ is the mean within-cluster effective sample size $\big(\sum_j \omega_{ij}\big)^2 / \sum_j \omega_{ij}^2$ of the tilt weights. Let $\bar{S}(\gamma) \coloneqq \frac{1}{N_{\mathrm{cov}}}\sum_{i\in\mathcal{C}}\sum_j \omega_{ij}(\gamma)\, S(\tau_{ij})$ be the tilted mean verified-step fraction of the corpus.

1. *Floor rule (targets the bound).* For a coverage budget u_0 and margin ρ, choose the largest floor meeting the budget: $\hat{\eta} \coloneqq \max\{\eta \in \mathcal{H} : \hat{u}(\eta) \leq u_0\}$ over a finite grid $\mathcal{H} \subset (\rho, 1]$.
2. *Tilt heuristic (not derived from the bound).* For an ESS budget ESS_0, choose $\hat{\gamma} \coloneqq \arg\max\{\bar{S}(\gamma) : \gamma \in \Gamma,\ \mathrm{ESS}(\gamma) \geq \mathrm{ESS}_0\}$ over a finite grid Γ.

All quantities involved are computable from $\{S(\tau)\}$ alone.

Proposition 5 (What the label-free rules do and do not target)**.** (a) *For every η, $\hat{u}(\eta)$ is an unbiased estimator of $u(\eta) = \mathbb{E}_x[(1-p_\phi(x;\eta))^K]$—exactly, not up to $o(1)$—and $N_{\mathrm{cov}}(\eta) = N(1-\hat{u}(\eta))$; neither depends on γ.* (b) *Among floors on the grid $\mathcal{H}$, the floor rule selects the largest η whose coverage terms $\big(u(\eta),\ N_{\mathrm{cov}}(\eta)^{-1/4}\big)$ are within the budget; since δ's explicit η-dependence, $\varepsilon_v(x)/(\eta-\rho)$, is decreasing in η, the rule minimizes the label-free-observable part of the bound's η-dependence subject to the coverage budget, up to the estimation error of $\hat{u}$ (binomial, $O(N^{-1/2})$).* (c) *The tilt heuristic does not minimize any term of Proposition 4: the bound's* explicit *γ-dependent terms are non-decreasing in γ (the implicit one, $\varepsilon_{\mathrm{apx}}$, is not sign-determined), so nothing in the bound recommends $\gamma > 0$. What $\bar{S}(\gamma)$ measures is the corpus's mean verified-step fraction, a label-free proxy for the per-trace soundness that Proposition 2 charges through $\varepsilon_v/(\eta-\rho)$ but whose* benefit *the Markov step discards. Agreement between $\hat{\gamma}$ and the validation-label oracle (Table 8) is therefore an empirical observation about this proxy, not a consequence of the theory.*

To state the practical claim of §2.2 precisely: the certification floor is selected by a rule that targets the same coverage quantities the finite-sample theory identifies, whereas the tilt is selected by a label-free heuristic whose justification is empirical. We keep the label-selected variants of both as oracles in the ablations.

A.7 The cascade inherits the bound

Let the deployed system retain the student's output when $\hat{v}_\psi \geq \kappa$ (retain rate r_κ) and escalate otherwise to the agentic teacher. Let $p^*(x, \hat{y}, \hat{g}) \coloneqq \mathbb{P}[\mathrm{Ver} = 1 \mid x, \hat{y}, \hat{g}]$ be the probability that the teacher's verifier certifies the student's typed output, and $Y \in \{0, 1\}$ the verifier's realized decision on a relabeled sample. Say $\hat{v}_\psi$ is *ξ-calibrated in ℓ_1* if $\mathbb{E}|\hat{v}_\psi - p^*| \leq \xi$; the absolute value is essential (a signed condition would allow cancellation and does not support the proof). Two facts govern how

ξ is reported. First, ξ is *not* directly identifiable from binary relabels, because p^* is a conditional probability and only Y is observed; what is observed is sandwiched around it:

$$\mathrm{ECE}(\hat{v}_\psi) \;\le\; \mathbb{E}\big|\hat{v}_\psi - p^*\big| \;\le\; \hat{\xi} \;\coloneqq\; \mathbb{E}\big|\hat{v}_\psi - Y\big|,$$

where the left inequality is Jensen's inequality applied within calibration bins and the right one is Jensen's inequality applied to $p^* = \mathbb{E}[Y \mid x, \hat{y}, \hat{g}]$. Reporting ECE alone therefore understates the term the bound needs; we report $\hat{\xi}$, which is measurable by re-running the teacher's verifier on student outputs (§2.4), and plug it into the bound, which is monotone in ξ. Second, if the verifier is deterministic given the typed output—which holds when the perception calls it issues are deterministic—then $p^* \in \{0, 1\}$ and $\hat{\xi}$ equals ξ exactly.

Assumption 3 (Student-side verifier primitives, on retained sets)**.** Assumption 1 holds for the distribution of *student* outputs under the teacher's verifier, restricted to any retained set $\{\hat{v}_\psi \ge \kappa\}$, with student-specific primitives $(\varepsilon^{\mathrm{s}}_v(x), \varepsilon^{\mathrm{s}}_{\mathrm{loc}}(x))$ and yields $(g_{\mathrm{s}}(x), \bar{q}_{\mathrm{s}}(x))$ whose implied bound δ_{s} (Proposition 2 instantiated on retained student outputs) does not depend on κ. All quantities are measured by re-running the verifier on student outputs over the annotated instrumentation subset (the relabeling pass of §2.4 provides exactly this data), at each reported operating point. This is a genuinely separate assumption, not a corollary of Assumption 1: the student's evidence distribution—its span-boundary error profile and its potential semantic shortcuts—differs from the teacher's, so teacher-side estimates do not transfer; and restricting to retained sets is necessary because the routing signal $\hat{v}_\psi$ could in principle concentrate poorly localized certified outputs among those it retains. The κ-independence of δ_{s} is what a single number in Proposition 6 requires; if the measured $\delta_{\mathrm{s}}(\kappa)$ varies, its maximum over the reported operating points is used. Appendix C reports the student-side diagnostic; on NExT-GQA the plug-in clips at 1, so it is valid but vacuous.

Proposition 6 (Cascade bound)**.** *Let $\hat{v}_\psi$ be ξ-calibrated in ℓ_1 against p^*, let δ_{s} be as in Assumption 3, and let $G^{\mathrm{esc}}_{\mathrm{T}}(\kappa) \coloneqq \mathbb{P}[\hat{y}_{\mathrm{T}} = y^\star \wedge \neg\mathrm{loc} \mid \hat{v}_\psi < \kappa]$ be the agentic teacher's grounding gap on the escalated subset. Then*

$$G(\text{cascade}) \;\le\; B(\kappa) \;\coloneqq\; r_\kappa\big((1-\kappa) + \delta_{\mathrm{s}}\big) \;+\; \xi \;+\; (1 - r_\kappa)\, G^{\mathrm{esc}}_{\mathrm{T}}(\kappa),$$

and the expected per-input cost is

$$c_{\mathrm{cascade}}(\kappa) \;=\; c_{\mathrm{s}} \;+\; (1 - r_\kappa)\, c_{\mathrm{T}}, \qquad c_{\mathrm{s}} \ll c_{\mathrm{T}},$$

not $r_\kappa c_{\mathrm{s}} + (1 - r_\kappa) c_{\mathrm{T}}$*: the student runs on* every *input, because its score is what decides whether to escalate, and c_{T} is the* additional *teacher cost on escalated inputs. At $c_{\mathrm{s}} = 0.05\, c_{\mathrm{T}}$ and 20% escalation this is $0.25\times$, not $0.24\times$; at $\kappa \to 1$ it is $1.05\times$, i.e. the cascade is strictly more expensive than the teacher when it escalates everything. Our student and teacher share no computation (different backbones, independently sampled frames), so no reuse credit is taken; a deployment that shared a frozen vision trunk could subtract its cost once, and would have to quantify the shared fraction. The grounding bound remains valid with ξ replaced by the measurable $\hat{\xi}$.*

Remark 3 (The cascade bound does not depend on distillation fidelity). Proposition 6 contains no ε: retained outputs are vouched for by their *predicted verification*, not by how well the student imitates the teacher. Even a poorly distilled student yields a system whose grounding is bounded—it simply escalates more often and costs more. Distillation quality (Propositions 3 and 4) governs *where on the cost axis* the system sits; verification governs the bound on its grounding.

Remark 4 (Where the bound is loose, and what it can and cannot certify). The term $r_\kappa(1-\kappa)$ charges, for every retained output, the full probability that the verifier would *reject* it, about which the bound says nothing. The student-side soundness plug-in clips at $\delta_{\mathrm{s}} = 1$. Consequently the bound exceeds 1 through 50% escalation and falls below 1 only at the very high-escalation operating point (about 80% in Table 16). As $\kappa \to 1$ the bound tends to $\hat{\xi} + G^{\mathrm{esc}}_{\mathrm{T}}$: in that limit its content is that the system's gap is at most $\hat{\xi}$ larger than the escalation target's on the inputs it escalates. This is the right shape for a cascade bound, but it means the bound can certify a *small absolute* gap only when the escalation target has one. On NExT-GQA the agentic teacher's own gap is ≈ 0.40, so no bound of this form is tight there; with a certified escalation target (e.g., best-of-K verifier-reranked teacher, whose gap is δ rather than $G(\pi_{\mathrm{T}})$) the same bound becomes $r_\kappa(1-\kappa) + \hat{\xi} + O(\delta)$, which is informative in the selective-prediction regime. The empirical improvement of the cascade at 10–20% escalation (Table 5) therefore comes from a mechanism the bound does not model—the correlation between low $\hat{v}_\psi$ and the student's *own* errors, visible in Table 16 as the 4–5 point margin of the teacher over the student on the escalated subsets at those operating points—rather than from the bound being

tight there. We state this rather than describe the bound as non-vacuous at the operating points we deploy.

Absent Assumption 3, the retained-set portion of the bound degrades gracefully but honestly to a *verifier-agreement* statement: retained outputs would, up to calibration error ξ, be certified by the teacher's verifier with probability at least κ, but nothing would connect certification to gold-span localization. The δ_{s} term is precisely what upgrades agreement into grounding, which is why we *measure* the student-side primitives on re-verified student outputs rather than assuming that teacher-trace soundness transfers. As $\kappa \to 1$ the system approaches the teacher's grounding at the teacher's cost.

A.8 A CERTIFIED, COST-OPTIMAL OPERATING POINT FOR THE CASCADE

Proposition 6 makes κ's effect on the bound explicit but does not by itself say how to *choose* κ. We give a selection rule in two parts: an assumption-free procedure that returns the cost-minimal routing policy meeting a target level of the *estimated* bound, and an additional checkable condition under which the bound is monotone and the procedure reduces to bisection. Two facts about r_κ shape the statement. On a finite calibration set of size n, $\hat{v}_\psi$ takes finitely many values $s_{(1)} < \cdots < s_{(n')}$, so $r_\kappa = \mathbb{P}[\hat{v}_\psi \geq \kappa]$ is a step function: all thresholds in $(s_{(k-1)}, s_{(k)}]$ induce the *same* routing decision and cost. And with the $\geq$ convention, r_κ is *left*-continuous, not right-continuous ($\{\hat{v}_\psi \geq \kappa\} \downarrow \{\hat{v}_\psi \geq \kappa_0\}$ as $\kappa \uparrow \kappa_0$, but $\uparrow \{\hat{v}_\psi > \kappa_0\}$ as $\kappa \downarrow \kappa_0$), so a feasible set defined by $B(\kappa) \leq \tau_G$ need not contain its infimum. Neither fact is a defect of the cascade; both mean that the object to select is a *routing policy*, not a real number.

Definition 2 (Routing-equivalence classes)**.** Two thresholds are routing-equivalent if they retain the same set of calibration outputs. The classes are the intervals $I_k = (s_{(k-1)}, s_{(k)}]$ for $k = 1, \ldots, n'$ (with $s_{(0)} := -\infty$) and $I_{n'+1} = (s_{(n')}, 1]$. The retain rate r_κ, the escalated set, $G_{\mathrm{T}}^{\mathrm{esc}}$, and the expected cost are constant on each class, so cost_k is well defined. *The bound $B(\kappa)$, however, is not:* it depends on κ explicitly through the term $r_\kappa(1-\kappa)$, which varies across a class even when routing does not. Assigning one bound to a whole class therefore requires a designated representative, and we take the class's *right endpoint* $\kappa_k := s_{(k)}$ (with $\kappa_{n'+1} := 1$), which is legitimate because every retained output satisfies $\hat{v}_\psi \geq s_{(k)}$ on that class, and optimal because it makes $1-\kappa$ smallest and hence B tightest among class members. We write $B_k := B(\kappa_k)$. A sharper alternative is to replace $r_\kappa(1-\kappa)$ by the exactly computable

$$m_k \;:=\; \mathbb{E}\big[\mathbf{1}\{\hat{v}_\psi \geq \kappa_k\}\,(1-\hat{v}_\psi)\big] \;\leq\; r_{\kappa_k}(1-\kappa_k),$$

which is genuinely constant on the class and needs no representative at all; the proof of Proposition 6 bounds $\mathbb{P}[R \wedge \neg C_v]$ by $m_k + \xi$ before any relaxation to $r_\kappa(1-\kappa)$, and m_k is non-increasing in k, so part (b) below is unaffected.

Assumption 4 (Relative student soundness, escalated-subset form)**.** Let $G_{\mathrm{T}}(s) := \mathbb{P}[\hat{y}_{\mathrm{T}} = y^\star \wedge \neg\mathrm{loc} \mid \hat{v}_\psi = s]$ be the agentic teacher's grounding gap conditioned on the self-verifier's score. Assume $\delta_{\mathrm{s}} \geq \sup_s G_{\mathrm{T}}(s)$ over the score range considered, i.e. the student-side soundness term on certified outputs is no smaller than the teacher's grounding gap on *any* routing slice (on a finite calibration set, on any routing-equivalence class); this implies $\delta_{\mathrm{s}} \geq G_{\mathrm{T}}^{\mathrm{esc}}(\kappa)$ for every κ. *This condition is checkable and must be checked rather than presumed.* On NExT-GQA, the student-side plug-in clips at $\delta_{\mathrm{s}} = 1$. The ordering condition therefore holds trivially because every grounding gap is at most 1, but this makes the condition uninformative as an empirical diagnostic. The deployed teacher's global grounding gap is $G(\pi_{\mathrm{T}}\text{-agentic}) = 0.822 - 0.423 \approx 0.40$, computed *within* the NExT-GQA population from that dataset's own Acc@QA and Acc@GQA columns of Table 1. We therefore use the assumption-free enumeration in part (a) of Proposition 7 for the reported operating points; part (b) remains a structural monotonicity result, but its condition is satisfied here only because the student-side bound is vacuous.

Proposition 7 (Bound-minimal cost-optimal operating point)**.** *Let $B(\kappa)$ be the right-hand side of Proposition 6, evaluated with measured r_κ, $\hat{\xi}$, δ_{s}, and $G_{\mathrm{T}}^{\mathrm{esc}}(\kappa)$ on the calibration set, and let τ_G be a target.* (a) [no monotonicity needed] *The feasible classes $\mathcal{K}_{\tau_G} := \{k : B_k \leq \tau_G\}$ are computable by one pass over the $n'+1$ routing-equivalence classes from the already-collected relabeling data, and whenever $\mathcal{K}_{\tau_G} \neq \emptyset$ the cost-minimal feasible policy is the feasible class with the largest retain rate, $k^\star := \arg\max_{k \in \mathcal{K}_{\tau_G}} r_k$ (ties in r_k are ties in cost and are broken toward the smaller B_k); any $\kappa^\star \in I_{k^\star}$ implements it, and all such $\kappa^\star$ are equivalent. The procedure costs $O(n \log n)$ (sorting the*

scores) and requires neither retraining nor any assumption on the shape of B. (b) [under Assumption 4] *B is non-increasing in κ; hence $\mathcal{K}_{\tau_G}$ is an upper set of classes $\{k_0, \dots, n'+1\}$, $k^\star = k_0$, and k_0 is found by bisection over the sorted score grid in $O(\log n)$ evaluations of B. In neither case is $\kappa^\star$ unique as a real number; it is unique as a routing policy.*

Remark 5 (This is a bound diagnostic, not a finite-sample certificate). Proposition 7 is a deterministic statement about the *plug-in* values B_k. It does not certify deployment risk, and we do not describe it as doing so. Three gaps separate it from a certificate, and all three are ours to disclose rather than to argue away. (i) *Estimation.* $\hat{\xi}$, δ_{s}, and $G_{\mathrm{T}}^{\mathrm{esc}}$ are estimates; δ_{s} in particular is a plug-in over per-input ratios estimated from $K = 8$ rollouts, so probabilities conditioned on rare trace classes are sparsely estimated, and averaging clipped ratios reduces variance without removing estimation bias. Percentile bootstrap intervals over inputs describe sampling variability; they are not conservative coverage statements for a ratio functional with a binding clip. (ii) *Selection.* $k^\star$ is chosen using the same calibration data that produced B_k, so $B_{k^\star}$ is a selected minimum and is optimistically biased. A genuine certificate would need simultaneous control across the $n'+1$ candidate classes—a union bound over classes combined with a one-sided interval for each term, which inflates the required target by roughly $\sqrt{\log(n'+1)/n}$ per term—and we do not currently supply it. (iii) *Data hygiene.* We do separate the four uses of data, and this separation is a precondition for anything stronger: the head is fit on one split, temperature-scaled on a second, the threshold selected on a third, and the reported numbers computed once on the test split. We accordingly call B_k an *empirical bound diagnostic*, and Remark 4 shows it is in any case vacuous at the operating points we deploy.

Two remarks on scope. First, Assumption 4 is weaker than requiring δ_{s} to be small in absolute terms; it only orders δ_{s} against the escalation target's own gap. But it is not the direction one should expect by default: an agentic teacher deployed with a single rollout is itself far from grounded on hard inputs, and the ordering can only hold when the escalation target is itself certified (e.g., verifier-reranked). Second, if the ordering holds on some slices and not others, part (b) applies per slice and part (a) globally; Table 16 reports the check globally and by question type. This result converts κ from a knob swept for a curve into an operating point selected by a stated procedure, subject to Remark 5: what the procedure returns is the cost-minimal policy meeting a target level of an *estimated bound*, not a finite-sample guarantee about deployment risk.

A.9 The bias of filtering, debiased

$\mathcal{R}_{\eta,\gamma}$ reweights the input distribution whenever verified yield correlates with difficulty. The inverse-yield weights in Eq. (5) are the self-normalized importance weights for this shift: in §A.11 we show the weighted empirical objective is a consistent estimator of the distillation risk under the *original* $\mathcal{D}$, at the usual variance cost governed by $\max_x 1/h(\hat{p}_\phi(x))$ (we clip weights and report the effective sample size). Under train–test input shift an additional TV term between input distributions must be added to Proposition 3.

A.10 Representation alignment and the self-verifier's calibration

The representation-alignment objective $\mathcal{L}_{\mathrm{align}}$ (Eq. (7)) is motivated in §2.3 as encouraging the student's evidence representations to carry belief-state structure. We now make precise what it implies for the self-verification head of §2.4. The result has two parts of very different status: a lemma converting the contrastive loss into a bound on representation distance, which is a short calculation under one stated normalization; and a proposition converting representation distance into calibration error, which requires an idealization (an unobservable "true belief encoding") and is offered as a conditional, testable hypothesis rather than as a guarantee.

Lemma 2 (InfoNCE controls distance to the positive under unit normalization)**.** *Let z and e_+ be ℓ_2-normalized, let the M negatives $\{e_b\}_{b\in\mathcal{B}^-}$ be ℓ_2-normalized, and let $\mathcal{L}_k = -\langle z, e_+\rangle/\iota + \log\big(e^{\langle z,e_+\rangle/\iota} + \sum_{b\in\mathcal{B}^-} e^{\langle z,e_b\rangle/\iota}\big)$ be the per-anchor term of Eq.* (7)*. Then*

$$\|z - e_+\|^2 \;\le\; 2\iota\,\mathcal{L}_k \;+\; 4 \;-\; 2\iota\log M.$$

Consequently, with $a(\theta) := \mathbb{E}_{(x,\tau),k}\|z_{\theta,k}(x) - e(b_{t(k)})\|$,

$$a(\theta) \;\le\; \sqrt{\big(2\iota\,\mathcal{L}_{\mathrm{align}}(\theta) + 4 - 2\iota\log M\big)_+} \;=:\; \Phi\big(\mathcal{L}_{\mathrm{align}}(\theta)\big).$$

The dependence is a square root, not an affine relation; the constants are explicit and depend only on the temperature ι and the number of negatives M, and no assumption is made about the negative distribution beyond normalization.

Assumption 5 (Lipschitz self-verification head and an idealized belief encoding)**.** Write $\hat{v}_\psi(x, \hat{y}, \hat{g}) = \sigma(f_\psi(z_\theta(x)))$ where $z_\theta(x)$ is the student's pooled evidence representation and f_ψ is the trained head. Assume f_ψ is L_{sv}-Lipschitz in z under the Euclidean metric on the unit sphere. Assume further that there exists an idealized "true belief encoding" $e^\star(x, k)$—a function of the gold event structure at evidence span k, not of the student's parameters—and a function g such that the verifier's certification probability satisfies $\mathbb{E}\big|p^*(x, \hat{y}, \hat{g}) - g(\{e^\star(x, k)\}_k)\big| \le \xi_0$, where ξ_0 collects the irreducible calibration error (annotation noise, temperature-scaling error, and the part of p^* that no representation of the gold event structure determines). Assume *separately* a **head-fit residual at the idealized encoding**: the trained-and-recalibrated head, evaluated at $e^\star$ rather than at the student's representation, approximates g, i.e.

$$\mathbb{E}\big|\sigma\big(f_\psi(\{e^\star(x, k)\}_k)\big) - g(\{e^\star(x, k)\}_k)\big| \;\le\; \xi_1.$$

This is a genuinely additional hypothesis: the mere *existence* of a function g of the idealized encoding that tracks p^* says nothing about whether the particular head we trained realizes it, so ξ_1 cannot be absorbed into ξ_0. Like ξ_0 it is not measurable, since $e^\star$ is an idealization. Define the *encoder discrepancy* $\Delta_e := \mathbb{E}_{(x,\tau),k}\|e(b_{t(k)}) - e^\star(x, k)\|$, the distance between the frozen encoding of the teacher's stated belief and the idealized encoding of what actually happened in the video.

Proposition 8 (Alignment and calibration, conditional form)**.** *Under Assumption 5, the ℓ_1 calibration error of the trained-and-recalibrated self-verification head satisfies*

$$\xi \;\le\; \xi_0 \;+\; \xi_1 \;+\; L_{\mathrm{sv}}\Big(\Phi\big(\mathcal{L}_{\mathrm{align}}(\theta)\big) + \Delta_e\Big),$$

with Φ from Lemma 2. Of the four terms, only $\Phi(\mathcal{L}_{\mathrm{align}})$ is measurable.

What the proposition does and does not say. The term $\Phi(\mathcal{L}_{\mathrm{align}})$ is measurable and decreases as alignment is minimized. The term Δ_e is *not* measurable—$e^\star$ is an idealization—and we deliberately keep it outside ξ_0 rather than folding it into an unrestricted residual: it is the quantity that would make the bound vacuous if the teacher's stated beliefs encode something other than the gold event structure, and hiding it would misrepresent the result. This proposition provides only an upper-bound characterization. Proposition 8 bounds ξ from *above*; a decreasing upper bound neither implies that the measured error decreases nor that a floor exists, since the true ξ may sit anywhere beneath the ceiling and may move in either direction. Only one direction is testable: if the measured $\hat{\xi}(\lambda_r)$ ever *exceeded* the evaluated right-hand side, the assumption set would be refuted—but the right-hand side contains three unmeasurable terms, so in practice not even that is checkable. We therefore report the sweep of $\hat{\xi}$ against λ_r (holding stages (i)/(ii) of §2.4 fixed) as a descriptive experiment, motivated by the bound and consistent with it, and we do not describe a co-movement of $\hat{\xi}$ and $\mathcal{L}_{\mathrm{align}}$ as confirming the proposition, nor its absence as refuting it. Table 18 reports $\hat{\xi}(\lambda_r)$ and $\mathcal{L}_{\mathrm{align}}(\lambda_r)$ with bootstrap intervals. If $\hat{\xi}(\lambda_r)$ does not track $\mathcal{L}_{\mathrm{align}}(\lambda_r)$, Assumption 5—most likely the existence of a single $e^\star$ sufficient for p^*, or a large Δ_e—should be treated as violated for this teacher/student pair. This is independent of both QA accuracy and the probing analysis of §2.3, which tests representation structure rather than calibration and cannot by itself validate the causal chain assumed here.

A.11 Full Proofs

Proof of Proposition 1. Substitute directly into Eq. (4). At $\eta = 0, \gamma = 0$: $\mathbf{1}[S(\tau) \ge 0] \equiv 1$ since $S(\tau) \in [0, 1]$ by definition, and $e^0 \equiv 1$, so $\mathcal{R}_{0,0}(\pi_{\mathrm{T}}) = \pi_{\mathrm{T}}$, i.e., $(\phi, \omega) \equiv (1, 1)$. For case 2, replace the retention score S by $S_{\mathrm{corr}}(\tau) = \mathbf{1}[\hat{y} = y^\star] \in \{0, 1\}$ and set $\eta = 1$: $\mathbf{1}[S_{\mathrm{corr}}(\tau) \ge 1] = \mathbf{1}[\hat{y} = y^\star]$, and at $\gamma = 0$ the tilt factor is again 1, giving $\phi = \mathbf{1}[\hat{y} = y^\star]$, $\omega \equiv 1$. For case 3, take $\eta \in (0, 1]$ at the operator's native score $S(\tau)$ and $\gamma = 0$: $\mathbf{1}[S(\tau) \ge \eta] = \mathrm{Ver}(\tau)$ by the definition of Ver in Eq. (1), and the tilt factor is again 1, giving $\phi = \mathrm{Ver}(\tau)$, $\omega \equiv 1$. In all three cases $\omega \equiv 1$ because $\gamma = 0$; no case uses $\gamma \ne 0$, which is exactly the family's remaining degree of freedom, establishing the claimed boundary-case characterization. □

Lemma 3 (Bounded functionals)**.** *For any f with range $[0, 1]$ and distributions μ, ν on a common space, $|\mathbb{E}_\mu f - \mathbb{E}_\nu f| \le \mathrm{TV}(\mu, \nu)$. In particular, since $G(\pi) = \mathrm{Correct}(\pi) - \mathrm{Grounded}(\pi) = \mathbb{E}_x\mathbb{P}_\pi[\hat{y} = y^\star \wedge \neg\mathrm{loc}]$ is the expectation of the* single $\{0, 1\}$*-valued functional $f = \mathbf{1}[\hat{y} = y^\star \wedge \neg\mathrm{loc}]$,*

the gap satisfies $|G(\mu) - G(\nu)| \le \mathbb{E}_x \mathrm{TV}$*, with no factor of two: bounding* Correct *and* Grounded *separately and applying the triangle inequality is a needless loss.*

Proof. Write $\mu - \nu = (\mu-\nu)_+ - (\mu-\nu)_-$; both parts have total mass $\mathrm{TV}(\mu,\nu)$. Then $|\mathbb{E}_\mu f - \mathbb{E}_\nu f| = |\int f\, d(\mu-\nu)_+ - \int f\, d(\mu-\nu)_-| \le \max\{\int f\, d(\mu-\nu)_+, \int f\, d(\mu-\nu)_-\} \le \mathrm{TV}(\mu,\nu)$, using $0 \le f \le 1$. □

Proof of Proposition 2. Work conditionally on an input x with $\mathbb{P}_{\pi_\mathrm{T}}[\hat{y} = y^\star \wedge \mathrm{Ver} = 1 \mid x] > 0$ and on the event $\{\hat{y} = y^\star\}$; all probabilities below are under the teacher's trace distribution so conditioned, and the per-input primitives $\varepsilon_v(x), \varepsilon_{\mathrm{loc}}(x)$ of Assumption 1 are the only verifier properties used. The final statement follows by averaging the *per-input bounds* over x; at no point is an average of ratios replaced by a ratio of averages. Let $U = \{\hat{g} \text{ does not localize } g^\star\}$ and split $U = U_{\mathrm{hi}} \cup U_{\mathrm{lo}}$ by whether the true-support fraction exceeds ρ.

Low-support part. On U_{lo}, at most ρT steps are genuinely supported, so even if all of them pass their checks, the event $\{S(\tau) \ge \eta\}$ requires the fraction of *falsely accepted* steps to be at least $\eta - \rho$. Let $A(\tau)$ be that fraction. The expectation we need is *conditional* on U_{lo}, and the conditional form of (V1) is stated precisely to license it: each unsupported step's false-accept probability conditioned on $\{x,\, \tau \in U_{\mathrm{lo}}\}$ is at most $\varepsilon_v(x)$, so by linearity of expectation (no independence needed), $\mathbb{E}[A \mid U_{\mathrm{lo}}] \le \varepsilon_v(x)$; a merely marginal per-step bound would *not* suffice under this conditioning, since detector failures could concentrate on the selected class. Markov's inequality then gives $\mathbb{P}[\mathrm{Ver} = 1 \mid U_{\mathrm{lo}}] \le \varepsilon_v(x)/(\eta - \rho)$.

Assembling by Bayes. Under the filtered teacher, the correct-and-ungrounded mass on x is

$$\begin{aligned}\mathbb{P}[U \mid \mathrm{Ver} = 1] &= \frac{\mathbb{P}[\mathrm{Ver} = 1 \mid U_{\mathrm{lo}}]\mathbb{P}[U_{\mathrm{lo}}] + \mathbb{P}[\mathrm{Ver} = 1 \mid U_{\mathrm{hi}}]\mathbb{P}[U_{\mathrm{hi}}]}{\mathbb{P}[\mathrm{Ver} = 1]} \\ &\le \frac{\frac{\varepsilon_v(x)}{\eta-\rho}\,(1 - g_\mathrm{T}(x)) + 1 \cdot \varepsilon_{\mathrm{loc}}(x)(1 - g_\mathrm{T}(x))}{\bar{q}(x)} = \frac{1 - g_\mathrm{T}(x)}{\bar{q}(x)}\Big(\varepsilon_{\mathrm{loc}}(x) + \frac{\varepsilon_v(x)}{\eta - \rho}\Big),\end{aligned}$$

where we used $\mathbb{P}[U] = 1 - g_\mathrm{T}(x)$, $\mathbb{P}[U_{\mathrm{hi}}] = \varepsilon_{\mathrm{loc}}(x)(1 - g_\mathrm{T}(x))$ by the definition of $\varepsilon_{\mathrm{loc}}(x)$ in (V2), $\mathbb{P}[\mathrm{Ver} = 1 \mid U_{\mathrm{hi}}] \le 1$, and $\mathbb{P}[\mathrm{Ver} = 1] = \bar{q}(x) > 0$ (all conditioned on x and $\hat{y} = y^\star$). Since the left-hand side is a probability, the bound may be clipped at 1, giving $\mathbb{P}[U \mid \mathrm{Ver} = 1, \hat{y} = y^\star, x] \le \delta(x)$. Finally, the filtered teacher's law on x is π_T conditioned on $\mathrm{Ver} = 1$ within the input, so

$$\mathbb{P}_{\pi_\mathrm{T}^{\mathrm{Ver}}}[\hat{y} = y^\star \wedge U \mid x] = \mathbb{P}[U \mid \mathrm{Ver} = 1, \hat{y} = y^\star, x] \cdot \mathbb{P}[\hat{y} = y^\star \mid \mathrm{Ver} = 1, x] \le \delta(x).$$

Inputs with $\mathbb{P}_{\pi_\mathrm{T}}[\hat{y} = y^\star \wedge \mathrm{Ver} = 1 \mid x] = 0$ contribute 0 to the left-hand side and are assigned $\delta(x) = 0$. Averaging over x yields $G(\pi_\mathrm{T}^{\mathrm{Ver}}) \le \mathbb{E}_x[\delta(x)] = \delta$. The constant-primitive form follows by replacing $\varepsilon_v(x), \varepsilon_{\mathrm{loc}}(x)$ with their suprema before clipping. □

Proof of Corollary 2. For each x, the unclipped ratio $t(x) = \frac{1-g_\mathrm{T}(x)}{\bar{q}(x)}\big(\varepsilon_{\mathrm{loc}}(x) + \varepsilon_v(x)/(\eta - \rho)\big)$ is affine in $(\varepsilon_v(x), \varepsilon_{\mathrm{loc}}(x))$ (holding $\eta, \rho, g_\mathrm{T}, \bar{q}$ fixed under the perturbation, so only the verifier-quality primitives vary), and $\delta(x) = \min\{1, t(x)\}$ with $t \mapsto \min\{1, t\}$ 1-Lipschitz. Hence $|\delta(x) - \delta'(x)| \le |t(x) - t'(x)|$, which is the input-weighted middle expression; averaging over x and bounding each per-input difference by its supremum gives the outer expression. Teacher-agnosticism (Corollary 1) justifies evaluating the bound at perturbed primitives without re-deriving Proposition 2: the proof used no property of the *unperturbed* verifier beyond Assumption 1. □

Proof of Lemma 1. Let ν be the $\gamma = 0$ (rejection-filtered) trace distribution and $\nu' \propto \nu\, w$ with $w(\tau) = e^{\gamma S(\tau)}$. On $\mathrm{supp}(\nu)$ we have $S \in [\eta, 1]$, hence $w \in [e^{\gamma\eta}, e^\gamma]$. For any event E, $\nu'(E) = \frac{\int_E w\, d\nu}{\int w\, d\nu} \le \frac{e^\gamma\, \nu(E)}{e^{\gamma\eta} \cdot 1} = e^{\gamma(1-\eta)}\nu(E)$. Apply this per input to the event $\{\hat{y} = y^\star \wedge \neg\mathrm{localized}\}$ and average over x: $G(\mathcal{R}_{\eta,\gamma}(\pi_\mathrm{T})) \le e^{\gamma(1-\eta)} G(\mathcal{R}_{\eta,0}(\pi_\mathrm{T})) \le e^{\gamma(1-\eta)}\delta$ by Proposition 2. □

Proof of Proposition 3. Apply Lemma 3 to the single functional $f = \mathbf{1}[\hat{y} = y^\star \wedge \neg\mathrm{loc}]$, whose expectation under a model's output law is exactly that model's grounding gap, for each x, and average: $|G(\pi_\theta) - G(P^{\mathrm{Ver}})| \le \mathbb{E}_x \mathrm{TV}(\pi_\theta, P^{\mathrm{Ver}}) \le \varepsilon$. Combining with Lemma 1 (which bounds $G(P^{\mathrm{Ver}})$ by δ_γ, and by δ when $\gamma = 0$) gives the claim. □

Proof of Theorem 1. Fix $m \in (0,1)$ and $\eta \in (0,1]$. Take a point-mass input distribution on x with gold $(y^\star, g^\star)$, and a teacher emitting, with probability m, the trace $\tau_0 = (\hat{y} = y^\star, \hat{g} = g_0)$ with $g_0 \cap g^\star = \emptyset$ and $v_t = 0$ for all t (so $S(\tau_0) = 0$), and with probability $1 - m$ the trace $\tau_1 = (\hat{y} = y^\star, \hat{g} = g^\star)$ with $v_t = 1$ for all t (so $S(\tau_1) = 1$). Let the verifier be exact on this family: unsupported steps are never accepted ($\varepsilon_v = 0$), there are no boundary-imprecise ungrounded traces ($\varepsilon_{\mathrm{loc}} = 0$), and $\rho = 0 < \eta$; Assumption 1 holds and Proposition 2 gives $\delta = 0$. By construction $G(\pi_{\mathrm{T}}) = m$.

(i) $\mathcal{R}_{\eta,\gamma}(\pi_{\mathrm{T}})$ places all mass on τ_1 (only $\bar{S} = 1 \geq \eta$ survives, the commit check passing there by construction), so $G(\pi_{\mathrm{T}}^{\mathrm{Ver}}) = 0$, $\delta_\gamma = 0$, and Proposition 3 gives $G(\pi_\theta) \leq \varepsilon$ for any student with fidelity ε.

(ii) Let μ be any output law with $\mathrm{TV}(\mu, P_{\pi_{\mathrm{T}}}) \leq \varepsilon$. By the sharpened form of Lemma 3 applied to $f = \mathbf{1}[\hat{y} = y^\star \wedge \neg\mathrm{loc}]$, $|G(\mu) - G(\pi_{\mathrm{T}})| \leq \varepsilon$, hence $G(\mu) \geq m - \varepsilon$.

Suppose some bound $G(\pi_\theta) \leq f(\delta, \varepsilon)$ held for raw-trace distillation with f independent of $G(\pi_{\mathrm{T}})$. Apply it to the construction with $\varepsilon \to 0$: it forces $f(0,0) \geq m$ for every $m < 1$, so $f(0,0) \geq 1$ and the bound is vacuous. Hence no non-vacuous bound of this form exists. We stress the scope: the theorem shows that *imitating a raw trace distribution inherits its grounding gap*, which is why unfiltered CoT distillation admits no analogous guarantee. It does not show that verification is the only mechanism capable of closing the gap empirically, nor that the measured gains of §3 are attributable to verification alone—that attribution rests on the matched-mechanism controls, not on this construction. □

Proof of Proposition 4. Write $\varepsilon(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\mathrm{TV}(\pi_\theta(\cdot \mid x), P^{\mathrm{Ver}}(\cdot \mid x))$ and split by the coverage event $C(x)$: $\varepsilon(\theta) \leq \mathbb{E}_x[\mathrm{TV} \cdot \mathbf{1}\{C(x)\}] + \mathbb{P}[\neg C] \leq \varepsilon_{\mathrm{cov}}(\theta) + u$, using $\mathrm{TV} \leq 1$ and $\mathbb{P}[\neg C(x)] = \mathbb{E}_x(1 - p_\phi(x))^K = u$.

Cluster-level risks. Fix the cross-fitting folds and condition on N_{cov}. Covered inputs are i.i.d. draws from $\mathcal{D}$ conditioned on coverage, because inputs are i.i.d. and $C(x_i)$ depends only on input i's own rollouts. For a covered input, let $\mu_\theta(x) := \mathbb{E}_{o\sim P^{\mathrm{Ver}}(\cdot|x)}\ell_\theta(x,o)$ and $\bar{\mu}_\theta(x) := \mathbb{E}[\bar{\ell}_\theta(x) \mid x, C(x)]$ be the population and cluster-conditional means of the loss; by the definition of $\varepsilon_{\mathrm{sn}}$, $\mathbb{E}_{x|C}[w^{\mathrm{HT}}(x)|\bar{\mu}_\theta(x) - \mu_\theta(x)|] \leq \varepsilon_{\mathrm{sn}}$ uniformly over $\theta \in \mathcal{F}$. The target risk is $L(\theta) = \mathbb{E}_{x|C}[w^{\mathrm{HT}}(x)\mu_\theta(x)]$.

Bias from estimated weights and from self-normalization. Within a fold, the cross-fitted $\hat{w}(x_i)$ is a function of x_i and of rollouts from *other* folds, hence independent of input i's own rollouts and coverage event given x_i; this is exactly the step where same-sample propensities would fail, since their bias conditional on retention neither vanishes nor is controlled by an unconditional moment. Therefore, conditionally on the other folds, $\mathbb{E}[\hat{w}(x_i)\bar{\ell}_\theta(x_i) \mid x_i, C(x_i)] = \hat{w}(x_i)\bar{\mu}_\theta(x_i)$, and

$$\sup_\theta \left|\mathbb{E}[\hat{L}(\theta)] - L(\theta)\right| \leq \sup_\theta \mathbb{E}_{x|C}\left|\hat{w}(x) - w^{\mathrm{HT}}(x)\right| \bar{\mu}_\theta(x) + \sup_\theta \mathbb{E}_{x|C} w^{\mathrm{HT}}(x)\left|\bar{\mu}_\theta(x) - \mu_\theta(x)\right| \leq B\varepsilon_w + \varepsilon_{\mathrm{sn}},$$

using $\bar{\mu}_\theta \in [0, B]$.

Concentration over clusters. Each summand $\hat{w}_i\bar{\ell}_\theta(x_i)$ is a function of one cluster (input i and its own rollouts) taking values in $[0, W_{\max}B]$; within a fold these summands are independent across i given the other folds. Hoeffding plus a union bound over $\mathcal{F}$ and over the F folds gives, with probability $1 - \zeta$, $\sup_\theta |\hat{L}(\theta) - \mathbb{E}\hat{L}(\theta)| \leq W_{\max}B\sqrt{\log(2F|\mathcal{F}|/\zeta)/(2N_{\mathrm{cov}}/F)} =: \Delta$; the dependence of traces *within* a cluster never enters, because the cluster is the sampling unit. ERM on $\hat{L}$ then yields $L(\hat{\theta}) \leq L(\theta^\star) + 2\Delta + 2B\varepsilon_w + 2\varepsilon_{\mathrm{sn}}$, i.e. $\mathbb{E}_{x|C}[w^{\mathrm{HT}}(x)\,\mathrm{KL}(P^{\mathrm{Ver}}\|\pi_{\hat{\theta}})] \leq \varepsilon_{\mathrm{apx}} + 2\Delta + 2B\varepsilon_w + 2\varepsilon_{\mathrm{sn}}$ (subtracting the weighted conditional entropy, common to all θ).

From KL to TV. The weights w^{HT} debias the covered input law back to $\mathcal{D}$ restricted to coverable inputs (Remark 6), so $\mathbb{E}_{x|C}[w^{\mathrm{HT}}(x)\,\mathrm{KL}(\cdot)] = \mathbb{E}_{x\sim\mathcal{D}}[\mathrm{KL}(\cdot) \mid C(x)]$ up to the mass already charged in u. By Pinsker and Jensen, $\varepsilon_{\mathrm{cov}}(\hat{\theta}) \leq \mathbb{E}_x\sqrt{\frac{1}{2}\mathrm{KL}} \leq \sqrt{\frac{1}{2}\mathbb{E}_x\mathrm{KL}} \leq \sqrt{\varepsilon_{\mathrm{apx}}/2} + \sqrt{\Delta} + \sqrt{B\varepsilon_w} + \sqrt{\varepsilon_{\mathrm{sn}}} + \sqrt{\varepsilon_Z}$, using subadditivity of the square root, and charging the deviation between the empirical and deterministic normalizers as the ε_Z term defined before the proposition. Substituting $\sqrt{\Delta} = (W_{\max}B)^{1/2}\big(F\log(2F|\mathcal{F}|/\zeta)/(2N_{\mathrm{cov}})\big)^{1/4}$ (the factor F is a constant absorbed into C)

and inserting $\varepsilon(\hat{\theta}) \le \varepsilon_{\mathrm{cov}} + u$ into Proposition 3—which contributes ε, not 2ε, by Lemma 3, and δ_γ rather than δ because the target is the tilted law—gives the display.

Coverage identities and γ-independence. $\mathbb{E}[N_{\mathrm{cov}}] = \sum_i \mathbb{P}[C(x_i)] = N\,\mathbb{E}_x[1-(1-p_\phi(x))^K]$; monotonicity from $\frac{d}{dp}(1-(1-p)^K) = K(1-p)^{K-1} \ge 0$ and $\frac{d}{dp}(1-p)^K \le 0$; $\mathbb{E}_x\mathbb{E}_{\tau\sim\pi_\phi}[\mathrm{Ver}(\tau)] = \mathbb{E}_x[p_\phi(x)]$ by definition of p_ϕ. Finally, $C(x) = \{\exists j : S(\tau^{(j)}) \ge \eta\}$ involves only the indicator in Eq. (4), not the tilt factor, so u and N_{cov} do not depend on γ.

The self-normalization term. At $\gamma = 0$, $\omega_{ij} = 1/m_i$ and, conditionally on the accepted index set, the accepted traces are i.i.d. from $\pi_{\mathrm{T}}(\cdot \mid x)$ restricted to $\{S \ge \eta\}$, which is $P^{\mathrm{Ver}}(\cdot \mid x)$; hence $\bar{\mu}_\theta = \mu_\theta$ and $\varepsilon_{\mathrm{sn}} = 0$. For $\gamma > 0$, $\bar{\ell}_\theta(x)$ is a self-normalized importance estimate of $\mu_\theta(x)$ with weights in $[e^{\gamma\eta}, e^{\gamma}]$; the standard bias bound for self-normalized estimators with bounded weight ratio $R = e^{\gamma(1-\eta)}$ and bounded integrand gives $|\bar{\mu}_\theta(x) - \mu_\theta(x)| \le R^2 B\,\mathbb{E}[1/m_i \mid x, C(x)]$, which is the expression quoted in §A.5. □

Proof of Proposition 5. (a) $\mathbf{1}[\max_j S(\tau_i^{(j)}) < \eta]$ is a Bernoulli variable with mean $\mathbb{P}[\neg C(x_i)] = \mathbb{E}_x[(1-p_\phi(x;\eta))^K] = u(\eta)$, where $p_\phi(x;\eta) = \mathbb{P}_{\tau\sim\pi_\phi(\cdot|x)}[S(\tau) \ge \eta]$; its sample mean $\hat{u}(\eta)$ is therefore unbiased with binomial variance $u(1-u)/N$, and $N_{\mathrm{cov}} = N(1-\hat{u})$ by definition. The event $\{S(\tau) \ge \eta\}$ does not involve γ, so neither quantity does. *(b)* Among $\eta \in \mathcal{H}$ with $\hat{u}(\eta) \le u_0$, the terms $u(\eta)$ and $N_{\mathrm{cov}}(\eta)^{-1/4}$ are within the budget up to the $O(N^{-1/2})$ estimation error of $\hat{u}$, while $\varepsilon_v(x)/(\eta-\rho)$ is strictly decreasing in η for every x; choosing the largest feasible η therefore minimizes the explicit η-dependence of δ subject to the coverage budget. (The primitives $\varepsilon_v(x), \varepsilon_{\mathrm{loc}}(x)$ are properties of the verifier on the fixed class U_{lo} and do not vary with η; ρ is fixed.) *(c)* By Lemma 1, the bound on $G(\mathcal{R}_{\eta,\gamma}(\pi_{\mathrm{T}}))$ scales as $e^{\gamma(1-\eta)}$, non-decreasing in γ; $\varepsilon_{\mathrm{sn}}$ is 0 at $\gamma = 0$ and non-negative otherwise; and the within-cluster ESS $(\sum_j \omega_{ij})^2 / \sum_j \omega_{ij}^2$ is maximized by uniform weights, i.e., at $\gamma = 0$. No term of Proposition 4 decreases in γ, so the bound's minimizer over γ is 0, and the tilt heuristic is not a consequence of the bound. □

Proof of Proposition 6. $G(\mathrm{cascade}) = \mathbb{P}[\hat{y}_{\mathrm{sys}} = y^\star \wedge \neg\mathrm{loc}]$; split on the routing event $R = \{\hat{v}_\psi \ge \kappa\}$ (retain, $\mathbb{P}[R] = r_\kappa$).

Escalated. On $\neg R$ the system outputs the agentic teacher's answer and evidence, contributing $\mathbb{P}[\neg R]\mathbb{P}[\hat{y}_{\mathrm{T}} = y^\star \wedge \neg\mathrm{loc} \mid \neg R] = (1-r_\kappa)\,G_{\mathrm{T}}^{\mathrm{esc}}(\kappa)$ exactly.

Retained. Because the student's evidence is typed, running the teacher's verifier on the output $(x, \hat{y}, \hat{g})$ is well defined; let C_v be the (auxiliary) certification event, with $\mathbb{P}[C_v \mid x, \hat{y}, \hat{g}] = p^*$. Then

$$\mathbb{P}[R \wedge \mathrm{correct} \wedge \neg\mathrm{loc}] \;\le\; \underbrace{\mathbb{P}[R \wedge \neg C_v]}_{\text{retained but would fail}} \;+\; \underbrace{\mathbb{P}[R \wedge C_v \wedge \mathrm{correct} \wedge \neg\mathrm{loc}]}_{\text{certified yet ungrounded}}.$$

For the first term, $\mathbb{P}[R \wedge \neg C_v] = \mathbb{E}[\mathbf{1}\{\hat{v}_\psi \ge \kappa\}(1-p^*)] \le \mathbb{E}[\mathbf{1}\{\hat{v}_\psi \ge \kappa\}(1-\hat{v}_\psi)] + \mathbb{E}|\hat{v}_\psi - p^*| \le (1-\kappa)r_\kappa + \xi$, using $1 - p^* \le (1-\hat{v}_\psi) + |\hat{v}_\psi - p^*|$ and ξ-calibration. For the second, write it as $r_\kappa\,\mathbb{P}[C_v \wedge \mathrm{correct} \wedge \neg\mathrm{loc} \mid R]$ and apply Proposition 2 to the distribution of *retained* student outputs under Assumption 3—the assumption supplies the student-side (V1)/(V2) *on the retained set*, which is what conditioning on R requires; without it this step is unavailable—giving $\mathbb{P}[C_v \wedge \mathrm{correct} \wedge \neg\mathrm{loc} \mid R] \le \delta_{\mathrm{s}}$, hence the term is at most $\delta_{\mathrm{s}} r_\kappa$. Summing the three contributions proves the bound. Replacing ξ by $\hat{\xi} \ge \xi$ (§A.7) preserves the inequality. For the cost identity, the student is evaluated on every input (its score is the routing statistic), so c_{s} is paid unconditionally, and the teacher's *additional* cost c_{T} is paid on the escalated fraction $1 - r_\kappa$. Note that no step used the distillation fidelity ε. □

Proof of Proposition 7. (a) On the calibration set, $R = \{\hat{v}_\psi \ge \kappa\}$ is the same set for all κ in a routing-equivalence class I_k, so r_κ, the escalated subset (hence $G_{\mathrm{T}}^{\mathrm{esc}}$), and the cost are constant on I_k. The relaxed bound $B(\kappa)$ is *not* constant on I_k because of its explicit $r_\kappa(1-\kappa)$ term; following Definition 2, evaluate it at the class right endpoint κ_k and write $B_k := B(\kappa_k)$, $r_k := r_{\kappa_k}$, $\mathrm{cost}_k = c_{\mathrm{T}} - r_k(c_{\mathrm{T}} - c_{\mathrm{s}})$. Each B_k is a finite expression in already-measured quantities, so all $n' + 1$ values are computable in one pass after sorting the scores. Cost is strictly decreasing in r_k because

$c_{\mathrm{T}} - c_{\mathrm{s}} > 0$; hence among feasible classes the one with the largest r_k is cost-minimal, and any $\kappa \in I_{k^\star}$ implements the same routing policy. No structure of B beyond finiteness was used.

(b) Fix $\kappa_1 < \kappa_2$. Since $\hat{v}_\psi$'s law does not depend on κ, $r_{\kappa_2} \le r_{\kappa_1}$. Write $B(\kappa) = r_\kappa(1-\kappa) + \xi + r_\kappa\big(\delta_{\mathrm{s}} - G_{\mathrm{T}}^{\mathrm{esc}}(\kappa)\big) + G_{\mathrm{T}}^{\mathrm{esc}}(\kappa)$. The term $r_\kappa(1-\kappa)$ is a product of two non-negative non-increasing functions, hence non-increasing ($r_{\kappa_1}(1-\kappa_1) \ge r_{\kappa_2}(1-\kappa_1) \ge r_{\kappa_2}(1-\kappa_2)$). For the remaining κ-dependent part, note that $(1-r_\kappa)G_{\mathrm{T}}^{\mathrm{esc}}(\kappa) = \mathbb{P}[\hat{v}_\psi < \kappa \wedge \hat{y}_{\mathrm{T}} = y^\star \wedge \neg\mathrm{loc}]$ is non-decreasing in κ (a larger escalated set), with increment $\mathbb{P}[\kappa_1 \le \hat{v}_\psi < \kappa_2 \wedge \hat{y}_{\mathrm{T}} = y^\star \wedge \neg\mathrm{loc}] = \mathbb{E}\big[\mathbf{1}\{\kappa_1 \le \hat{v}_\psi < \kappa_2\}\, G_{\mathrm{T}}(\hat{v}_\psi)\big] \le (r_{\kappa_1} - r_{\kappa_2}) \sup_s G_{\mathrm{T}}(s)$; meanwhile $r_\kappa\delta_{\mathrm{s}}$ decreases by $(r_{\kappa_1} - r_{\kappa_2})\delta_{\mathrm{s}}$. Under Assumption 4, $\delta_{\mathrm{s}} \ge \sup_s G_{\mathrm{T}}(s)$, so the decrease dominates the increase and $r_\kappa\delta_{\mathrm{s}} + (1-r_\kappa)G_{\mathrm{T}}^{\mathrm{esc}}(\kappa)$ is non-increasing; ξ is constant. Hence B is non-increasing in κ, so B_k is non-increasing in k and $\mathcal{K}_{\tau_G}$ is an upper set $\{k_0, \dots, n'+1\}$; the cost-minimal feasible class is k_0 by (a), and a monotone sequence's first feasible index is found by bisection in $O(\log n')$ evaluations. Nothing about continuity of r_κ in κ was used, which is why the left-continuity of r_κ (§A.8) is harmless: the object selected is a class, not a real threshold. □

Proof of Lemma 2. Write $s_+ = \langle z, e_+\rangle$ and $s_b = \langle z, e_b\rangle$; by normalization $s_b \ge -1$ for every b, so $\sum_{b\in\mathcal{B}^-} e^{s_b/\iota} \ge Me^{-1/\iota}$ and $\mathcal{L}_k \ge -s_+/\iota + \log\big(Me^{-1/\iota}\big) = -s_+/\iota + \log M - 1/\iota$, i.e. $s_+ \ge \iota\log M - 1 - \iota\mathcal{L}_k$. Since $\|z - e_+\|^2 = 2 - 2s_+$ on the unit sphere, $\|z-e_+\|^2 \le 2 - 2\iota\log M + 2 + 2\iota\mathcal{L}_k = 2\iota\mathcal{L}_k + 4 - 2\iota\log M$. Averaging over anchors gives $\mathbb{E}\|z-e_+\|^2 \le 2\iota\mathcal{L}_{\mathrm{align}} + 4 - 2\iota\log M$ (and the left-hand side is non-negative, whence the $(\cdot)_+$), and Jensen's inequality $\mathbb{E}\|z-e_+\| \le \sqrt{\mathbb{E}\|z-e_+\|^2}$ gives the bound on $a(\theta)$. □

Proof of Proposition 8. Let $p^\dagger := g(\{e^\star(x,k)\}_k)$, so that $\mathbb{E}|p^* - p^\dagger| \le \xi_0$ by Assumption 5. Let $\hat{v}^\star_\psi$ denote the head evaluated at the idealized encodings $e^\star$ in place of z_θ, and let $\hat{v}^e_\psi$ denote it evaluated at the frozen teacher encodings $e(b_{t(k)})$. By L_{sv}-Lipschitzness of f_ψ and 1-Lipschitzness of σ, $|\hat{v}_\psi - \hat{v}^e_\psi| \le L_{\mathrm{sv}}\|z_\theta - e(b)\|$ and $|\hat{v}^e_\psi - \hat{v}^\star_\psi| \le L_{\mathrm{sv}}\|e(b) - e^\star\|$ pointwise. Then

$$\mathbb{E}|\hat{v}_\psi - p^*| \;\le\; \mathbb{E}|\hat{v}_\psi - \hat{v}^e_\psi| + \mathbb{E}|\hat{v}^e_\psi - \hat{v}^\star_\psi| + \mathbb{E}|\hat{v}^\star_\psi - p^\dagger| + \mathbb{E}|p^\dagger - p^*| \;\le\; L_{\mathrm{sv}}\, a(\theta) + L_{\mathrm{sv}}\,\Delta_e + \xi_1 + \xi_0,$$

where the third term is bounded by the head-fit hypothesis of Assumption 5 rather than folded silently into ξ_0. Applying Lemma 2 to $a(\theta)$ gives the claim. The three unmeasurable quantities Δ_e, ξ_1, and ξ_0 are kept separate because they fail for different reasons—respectively, a belief encoder that does not track the gold event structure; a head that does not realize the idealized functional; and irreducible noise in p^*—and lumping them would hide which hypothesis a negative result falsifies. □

Remark 6 (Debiasing weights: unbiasedness, variance, and how ε_w is estimated). An input contributes to $\mathcal{D}_{\mathrm{ver}}$ with probability $h(p_\phi(x)) = 1 - (1 - p_\phi(x))^K$, so the covered empirical input law has density $\propto h(p_\phi(x))$ w.r.t. $\mathcal{D}$, and the ideal Horvitz–Thompson weights $w^{\mathrm{HT}}(x) \propto 1/h(p_\phi(x))$ (normalized to mean one over covered inputs) make the weighted loss unbiased for the $\mathcal{D}$-risk restricted to coverable inputs. Three implementation facts matter for validity. *(i) Same-sample propensities are invalid:* an estimate of $p_\phi(x)$ built from the same K rollouts that determine inclusion is statistically dependent on the inclusion event—among retained inputs it is necessarily positive and upward-biased for low-yield inputs—so $1/h(\hat{p}_\phi)$ built from it is not a Horvitz–Thompson weight and supports no unbiasedness claim. We therefore *cross-fit*: a yield predictor is trained on held-out folds of inputs and applied out-of-fold, making $\hat{p}_\phi(x)$ independent of the input's own rollouts. *(ii) The residual error is measured, not assumed away:* predictor error and clipping at $W_{\max}$ induce the bias $\varepsilon_w = \mathbb{E}_{x|C}|\hat{w} - w^{\mathrm{HT}}|$ that appears explicitly in Proposition 4. Because the exact $p_\phi(x)$ is unavailable, ε_w is estimated on a probe subset by replacing $p_\phi(x)$ with a high-budget Monte-Carlo yield ($K' = 64$ rollouts per probe input, whose binomial error is negligible relative to the predictor's), forming w^{HT} from it with the same mean-one normalization used for $\hat{w}$, and averaging $|\hat{w} - w^{\mathrm{HT}}|$ over covered probe inputs; the same probe rollouts are used to estimate $\varepsilon_{\mathrm{sn}}$ by comparing K-rollout cluster averages of the loss against their K'-rollout counterparts. Both estimates are reported alongside the clipped mass and the effective sample size. *(iii) Normalization after clipping:* clipping at $W_{\max}$ removes mass from high-weight inputs, so both $\hat{w}$ and w^{HT} are renormalized to mean one *after* clipping, and the discrepancy between the two normalizers is part of ε_w by construction. The same-sample variant is retained only as an ablation (Table 8), where its bias visibly costs grounded accuracy.

B CONTROLLER TRAINING DETAILS

Algorithm 1 VERA training (deliberate, certify, amortize)

Require: teacher with verifier; softened reference controller π_{ref}; corpus $\mathcal{X}$; budget K; dials η, γ

Stage 1 — Deliberate: learn the verification-guided controller

for each $x \in \mathcal{X}$ **do**

  sample K traces $\{\tau_i\} \sim \pi_\phi(\cdot \mid x)$; score $r(\tau_i)$ by Eq. (2); form pairs (τ^+, τ^-)

end for

update ϕ by minimizing $\mathcal{L}_{\text{ctrl}}$ (Eq. (3)); **backbones frozen**

Stage 2 — Certify: tilted verification filtering

roll out the controller-driven teacher; apply $\mathcal{R}_{\eta,\gamma}$ (Eq. (5)) to obtain $\mathcal{D}_{\text{ver}}$ with weights $w(\tau)$

retain the full labeled pool $\mathcal{D}_{\text{all}} = \{(x, \tau, \text{Ver}(\tau))\}$ ▷ rejected traces are self-verifier supervision

Stage 3 — Amortize the reasoning

train π_θ on $\mathcal{D}_{\text{ver}}$ by minimizing $\mathcal{L}_{\text{distill}} + \lambda_r \mathcal{L}_{\text{align}}$ (Eqs. (6), (7))

Stage 4 — Amortize the verifier

pre-train $\hat{v}_\psi$ on $\mathcal{D}_{\text{all}}$ by minimizing $\mathcal{L}_{\text{sv}}$ (Eq. (8))

relabel: run the teacher's verifier offline on student outputs over a held-out split; fine-tune and temperature-scale $\hat{v}_\psi$ on these labels ▷ calibrates on the deployment distribution

return single-pass student $(\pi_\theta, \hat{v}_\psi)$; optionally deploy cascade with threshold κ

Action repertoire. SELECT-SCRIPT(k), PERCEPTUAL-QUERY$(r, [t_1, t_2])$, REFLECT, COMMIT. **Softened reference and clipping.** π_{ref} replaces each hard threshold decision of the teacher's original controller by a tempered softmax over the same decision statistics (temperature T_{ref}), guaranteeing full support. T_{ref} is annealed only until π_{ref} matches the hard controller's argmax decisions with probability 0.9 (never fully deterministic), and per-step log-ratios $\log \frac{\pi_\phi(a_t|b_t)}{\pi_{\text{ref}}(a_t|b_t)}$ are clipped to $[-c, c]$ with $c = 5$ before summing, bounding the variance that a near-deterministic reference would otherwise reintroduce. **Reward.** Eq. (2); label-free ($\alpha = 0$, used for all main results) and label-augmented ($\alpha > 0$, ablation only) variants. **Matched distillation-signal controls.** Isolating the selection signal requires corpora matched on everything except selection, which filtering by construction disturbs. For every control in Table 3 we match (i) the number of retained (x, τ) examples, by subsampling the larger corpora to the size of the smallest; (ii) the number of covered inputs, so no control sees a different input distribution; (iii) the total target-token count; and (iv) the optimization budget in steps, batch size, and schedule. Unmatched variants are reported in Table 8. **Pair formation.** Default: top-vs-bottom (the top-ranked trace paired with each of the bottom $\lfloor K/2 \rfloor$); compared against all-pairs and adjacent-pairs at matched pair budgets (Table 9). Ties in $r(\cdot)$ are dropped. **Step-level variant.** Using v_t as a process signal, per-step advantages $A_t = v_t - \bar{v}$ weight a token-level policy-gradient update; compared against trajectory DPO in Table 9. **Coverage-sensitive variant.** Because the mean-yield objective need not raise p_ϕ pointwise (§A.5), we also train a variant that targets the uncovered mass $u = \mathbb{E}_x[(1 - p_\phi(x))^K]$ directly. The obvious construction does not work and we state why: $-(1 - \hat{p}_\phi(x))^K$ is an *input-level* quantity, identical for every trajectory sampled on x, so substituting it for $\text{Ver}(\tau)$ in Eq. (2) makes $r(\tau^+) = r(\tau^-)$ for all within-input pairs and the DPO objective of Eq. (3) receives no signal at all. We instead keep the trajectory ranking at $r(\tau)$ and move the coverage term into a *per-input weight* on that input's preference pairs,

$$\Lambda(x) \;=\; K\big(1 - \hat{p}_\phi(x)\big)^{K-1} \;=\; -\,\tfrac{\partial}{\partial p}\big(1-p\big)^K\Big|_{p=\hat{p}_\phi(x)}, \qquad \mathcal{L}^{\text{cov}}_{\text{ctrl}}(\phi) = \mathbb{E}_x\big[\Lambda(x)\,\mathcal{L}_{\text{ctrl}}(\phi; x)\big],$$

so that a unit of yield improvement is weighted by how much it reduces u at that input; this upweights near-zero-yield inputs, which are exactly the ones filtering silently drops, and is a valid gradient-reweighting of the same within-input ranking problem. $\hat{p}_\phi(x)$ is the cross-fitted estimate of §2.2, held fixed within an epoch so the weight is not differentiated through. A REINFORCE alternative, which treats $-(1 - \hat{p}_\phi(x))^K$ as an episode return with an input-level baseline and therefore does not need within-input ranking at all, is reported alongside. Table 9 compares the two against mean-yield (matched within noise on our distribution, consistent with Figure 3(e)). **Sampling and optimization.** K=8 traces per input, temperature 1.0; DPO β_{KL}=0.1; AdamW, lr 10^{-5}, 3,000 steps; backbones frozen.

Architecture and training, in full. *Students.* Gemma 3 instruction-tuned checkpoints at 1B, 4B, and 12B (Team et al., 2025); the 4B and 12B checkpoints are natively multimodal and use their own vision tower. *Vision front-end.* The 1B checkpoint is text-only, so it is paired with the same frozen

Table 9: **Controller-training variants** on NExT-GQA (4B student, three seeds; everything downstream of the controller fixed). Yield is the mean verified yield $\mathbb{E}_x[p_\phi(x)]$ at K=8; $u = \mathbb{E}_x[(1 - p_\phi(x))^K]$ is the uncovered mass.

Controller variant	Yield (%)	u	Acc@GQA
Hand-set thresholds (no learning)	34.0	0.130	39.1
Trajectory DPO, top-vs-bottom pairs (default)	46.3	0.043	41.0
all-pairs (matched pair budget)	45.6	0.048	40.8
adjacent-pairs (matched pair budget)	44.9	0.055	40.6
Step-level policy gradient ($A_t = v_t - \bar{v}$)	45.1	0.052	40.7
Coverage-sensitive weighting $\Lambda(x)$ (DPO)	45.8	0.036	41.1
Coverage REINFORCE ($-(1-\hat{p}_\phi)^K$ return)	45.2	0.039	41.0
Label-augmented reward (α=0.5, ablation only)	46.5	0.042	41.2

vision tower and a freshly initialized multimodal projector mapping visual tokens into the language embedding space; in all three students the projector is trained and the vision tower frozen during distillation.

Cost accounting. The amortization break-even is computed as the query volume $Q^\star$ at which $C_{\text{offline}} + Q\, c_{\text{s}} = Q\, c_{\text{T}}$, where C_{offline} sums trace collection (K rollouts per training input, including all perception calls), controller training, distillation, and self-verifier relabeling, and $c_{\text{s}}, c_{\text{T}}$ are per-query student and teacher costs measured under the matched protocol of §3. Table 10 reports the resulting $Q^\star$ by backbone scale and escalation rate; trace collection dominates C_{offline}, so halving K nearly halves the break-even at the coverage cost quantified in Figure 3(f) ($0.957 \to 0.860$ covered at K=8 $\to$ 4).

Label-use disclosure, in full. All training signals are label-free: the controller reward at $\alpha = 0$ (Eq. (2)), tilted filtering (Eq. (4)), distillation targets (Eq. (6)), and self-verifier labels (Eq. (8)) consume no gold answers or spans anywhere in the training loss. Three uses of annotation remain outside the training loss. *Model selection:* γ, η, κ, and the loss coefficients $\lambda_\cdot$ are selected on the NExT-GQA *validation* split using gold answers and spans; the test split is touched once, for final numbers. Table 8 additionally reports a fully verifier-selected variant—η by the coverage-budget floor rule and γ by the tilted mean verified-step fraction at matched effective sample size (Definition 1; the former is justified by Proposition 5, the latter is a heuristic), κ by target escalation rate, $\lambda_\cdot$ by validation loss on $\mathcal{D}_{\text{ver}}$—which matches label-based selection within seed noise (40.8 vs. 41.0 Acc@GQA). We therefore use "label-free" to describe the training and curation signals, while disclosing validation-time label use separately. *Measurement of the theory:* $\hat{\varepsilon}_v$, $\hat{\varepsilon}_{\text{loc}}$, $\hat{g}_{\text{T}}$, the student-side primitives of Assumption 3, and the direct gap $\hat{\delta}$ are estimated on a 1,000-clip annotated instrumentation subset drawn from the validation split. *Evaluation:* all reported metrics use gold answers and spans, as is standard. Annotations are never used in any training loss.

Table 10: **Break-even accounting** on NExT-GQA, in teacher-query equivalents, against the ordinary agentic teacher (1.0×). C_{offline}: 31,500 training inputs × K=8 rollouts = 2.52×10^5, + controller training (0.09×10^5) + distillation and self-verifier relabeling (0.06×10^5) = 2.67×10^5. $Q^\star = C_{\text{offline}}/(c_{\text{T}} - c_{\text{sys}})$ per unit teacher cost; cascade costs follow $c_{\text{s}} + (1 - r_\kappa)c_{\text{T}}$ (Proposition 6). The two zero-offline-cost alternatives are listed for comparison: they never break even because they have nothing to amortize, and each must be compared against the deployed system on its own terms.

Deployed system	C_{offline}	$c_{\text{sys}}/c_{\text{T}}$ (4B / 12B)	$Q^\star$ (4B)	$Q^\star$ (12B)
Student only	2.67×10^5	0.05 / 0.09	2.8×10^5	2.9×10^5
Cascade, 10% escalation	2.67×10^5	0.15 / 0.19	3.1×10^5	3.3×10^5
Cascade, 20% escalation	2.67×10^5	0.25 / 0.29	3.6×10^5	3.8×10^5
Verify-then-escalate (20%)	0	0.35 / 0.39	n/a	n/a
Test-time reranking (best-of-8)	0	8.0 / 8.0	n/a	n/a

C Additional Experimental Results

Amortization break-even (Q1). Amortization is only worth its offline bill at sufficient deployment volume, so we report the bill and the break-even explicitly rather than only per-query cost. Under the accounting of Appendix B, the offline cost on NExT-GQA totals 2.67×10^5 teacher-query equivalents (31,500 inputs × K=8 rollouts = 2.52×10^5, plus 0.15×10^5 for controller training, distillation, and relabeling), dominated ($\approx 94\%$) by trace collection and hence roughly linear in K.

Table 11: **Absolute cost and break-even sensitivity** on NExT-GQA, 4B student, A100 GPU. *Latency is single-query and unbatched for every system* (batch 1, one query in flight), which is the only setting in which the systems are directly comparable; *throughput* is measured separately at each system's own best batch size (student 16; the agentic loop is sequential and cannot exceed 1 concurrent query, so its throughput is a per-worker figure). The FLOP-and-call ratios of Tables 1 and 10 are the matched-protocol accounting and are hardware-independent; the wall-clock figures here are not, and the two should not be read as the same quantity. Latency is the mean per query (ms) over the test split, so the cascade and verify-then-escalate rows average the escalated and non-escalated queries, counting the student's forward pass on *every* query and adding the teacher only on the escalated fraction (no reuse; Prop. 6), consistent with the FLOP ratios of Tables 1 and 5; memory is peak GB; calls/frames are means per query. Right block: $Q^\star$ (student-only) recomputed under the stated change, all else fixed.

System	Latency (batch 1)	Throughput (own batch)	Memory	Perception calls	Frames
Student only	210	38	14	0	32
Verify-then-escalate	1,530	0.68	22	3.1	55
Cascade (20%)	1,170	1.0	22	2.3	51
Teacher (agentic)	4,800	0.21	22	11.4	96
Break-even sensitivity ($Q^\star$, student-only, $\times 10^5$)					
$K = 4$ / 8 / 16		1.5 / 2.8 / 5.5			
Perception calls per input $\times 0.5$ / $\times 1$ / $\times 2$		1.6 / 2.8 / 5.2			
Video length tercile (short / mid / long)		2.0 / 2.8 / 4.3			
Wall-clock instead of FLOP accounting		2			.7

The student-only deployment breaks even at $Q^\star \approx 2.8 \times 10^5$ queries (4B backbone; 2.9×10^5 at 12B) and the 20%-escalation cascade at $Q^\star \approx 3.6 \times 10^5$ (Table 10). Below these volumes the student does not pay for itself, but each alternative needs its own comparison, since $Q^\star$ above is computed against the *ordinary* teacher at $1.0\times$: best-of-8 reranking costs $8.0\times$ per query and so is never cost-competitive at any volume, whereas verify-then-escalate at $0.35\times$ has zero offline cost and matches our cascade within noise. The relevant crossover against *it* is therefore not $Q^\star$: at matched accuracy the cascade's per-query saving over verify-then-escalate is only $0.10\, c_{\mathrm{T}}$, so the offline bill is repaid against verify-then-escalate only above $C_{\mathrm{offline}}/(0.10\, c_{\mathrm{T}}) \approx 2.7 \times 10^6$ queries (the student-only deployment, at lower grounded accuracy, crosses it at $C_{\mathrm{offline}}/(0.30\, c_{\mathrm{T}}) \approx 8.9 \times 10^5$), which is what confines our cascade's advantage to the high-volume regime and makes verify-then-escalate the rational choice below those volumes. Two caveats keep this claim in scope: the "teacher-query equivalent" is a FLOP-and-call unit under our matched protocol, and its mapping to wall-clock time depends on hardware utilization, video decoding, and batching, which differ between a sequential, perception-bound agentic loop and a batchable forward pass (Table 11 therefore reports absolute latency, throughput, and memory); and because C_{offline} is dominated by trace collection, $Q^\star$ scales roughly linearly in K, in the per-input perception-call count, and with video length through both (sensitivities tabulated there), so the break-even is a scenario-dependent estimate.

Evaluation protocol and provenance. All tables use a single evaluation pipeline with dataset, split, population, metric, seed, and aggregation rule fixed. Subset metrics are evaluated directly on the corresponding examples, and the mixture identities are checked explicitly. Head fitting, temperature scaling, selection of κ and the remaining dials, and final evaluation use four disjoint splits; the test split is touched once. *Training-split provenance for Table 1:* the NExT-QA and NExT-GQA columns come from one student trained on the NExT-QA training split (NExT-GQA questions are a span-annotated subset of NExT-QA's evaluation questions), the IntentQA column from IntentQA's own split (zero-shot transfer: Table 7), and every in-pipeline EgoSchema entry is zero-shot transfer of the NExT-trained student, since EgoSchema has no training split.

Re-evaluated external baselines (‡). Every ‡ entry in Table 1 is measured by us with the official evaluator under the matched cost protocol of §3: the released checkpoint of each system, 32 frames unless the system prescribes its own budget, greedy decoding, and a fixed answer-plus-timestamps prompt template; the per-baseline prompts, decoding parameters, and checkpoint identifiers are listed in `baselines/` of the anonymized repository. Reproduced measurements are never mixed with originally published numbers in the same comparison; published entries carry no ‡.

Absolute costs and break-even sensitivity. The relative figures in Tables 1 and 10 are FLOP-and-call accountings under the matched protocol of §3: matched frame-sampling policy, resolution,

backbone scale, hardware, and decoding budget, with teacher-side perception calls counted individually. *Batching is deliberately not matched*, because it cannot be: the agentic teacher is a sequential, perception-bound loop that admits one query in flight, whereas the student is a batchable forward pass, so forcing a common batch size would penalize one system for an architectural property rather than measure it. We therefore report the two quantities separately. Single-query latency is reported at batch 1 for every system and is directly comparable. Throughput is reported at each system's own best batch size and is not; it is a deployment figure whose value depends on the concurrency the operator can supply. Table 11 reports both, plus peak memory, perception calls, decoded frames, and the sensitivity of $Q^\star$ to the three quantities that dominate C_{offline}. Two wall-clock entries deserve reconciliation with the FLOP accounting. Verify-then-escalate's latency implies a verifier share of $\approx 0.075\, c_{\text{T}}$, below its $0.10\, c_{\text{T}}$ FLOP share (Table 5), because the verifier's perception calls over already-cited spans parallelize better than the sequential agentic loop whose cost defines c_{T}; the FLOP figure is the hardware-independent one and is what the cost columns of Tables 1 and 5 use. And the zero-shot student's $0.04\times$ latency against its $0.05\times$ FLOPs reflects its shorter decode: prompted without typed-evidence or rationale targets it emits fewer output tokens, which reduces wall-clock slightly more than the (rounded) FLOP ratio.

Measuring the analysis. Figure 3(a–c) plots the acceptance-conditioned localization rates, the soundness diagnostic, and the cascade-bound terms of Tables 14–16; Figure 3(d–f) reports the controller and coverage diagnostics that remain independent of the soundness plug-in. Table 15 reports the aggregate localization diagnostic. The direct filtered-teacher correct-but-ungrounded mass is approximately 0.40, while the Proposition 2 plug-in saturates at the probability ceiling and is vacuous. Accordingly, we do not use this bound as an empirical certificate or as evidence for the benefit of moderate tilt.

Protocol for Table 2. Both are matched-budget replications, not new methods. For the second teacher: VideoMind's planner is left untouched and no controller is learned, so the replication covers the certify and amortize stages only; the "Teacher" cell of Table 2 is a single greedy agentic pass with the verifier role active under our matched 96-frame budget, whereas the vanilla reference row is VideoMind's released 32-frame evaluation setting without verifier re-query, which is why the two differ by 10.3 Acc@GQA with the planner untouched. Its verifier role supplies the per-step indicator v_t, so $\bar{S}$ and Ver (Eq. (1)) are computed exactly as in §2.2 but from a verifier we did not build; η and γ are re-selected by the label-free rule of Definition 1 rather than copied across, since the two verifiers' score distributions differ. The primitives $(\hat{\varepsilon}_v, \hat{\varepsilon}_{\text{loc}}, \hat{g}_{\text{T}}, \hat{\bar{q}})$ of Assumption 1 are remeasured on the instrumentation subset for the new teacher, which is what Corollary 1 requires and what makes the resulting δ comparable. For the new benchmarks: ReXTime is multiple-choice with separate moment localization, so Acc@GQA is computed by the same evaluator as NExT-GQA; CG-Bench is scored by its own released clue-grounded protocol, and we report its metric under its own name rather than folding it into an Acc@GQA column. Training corpora are collected on each benchmark's own training split; no cross-benchmark training is used, so these are three independent replications rather than a transfer result. For ReXTime and CG-Bench, (η, γ) are selected on each benchmark's own trace pool by the label-free rule of Definition 1 (the same rule used for the VideoMind replication; unlike the primary NExT-GQA run, no validation-label selection is used for the replications), and κ by target escalation rate, mirroring the NExT-GQA protocol. Table 2's cells are single replication runs; Table 12 reports paired-bootstrap 95% CIs over test questions for each verified-minus-correctness Δ.

Table 12: **Uncertainty for Table 2**: paired-bootstrap 95% CIs (10^4 resamples over test questions) for the verified $-$ correctness-filtered Δ on the joint metric of each setting.

Setting	Δ (joint metric)	95% CI
VideoMind-7B / NExT-GQA	$+4.1$	$[+2.9, +5.3]$
LINGUA / NExT-GQA	$+5.0$	$[+3.8, +6.2]$
LINGUA / ReXTime	$+4.2$	$[+2.7, +5.7]$
LINGUA / CG-Bench	$+3.6$	$[+1.8, +5.4]$

Seed statistics for Table 7 (scale rows). Each scale row of Table 7 is the mean over three seeds of the same pipeline re-run end-to-end at that student scale.

Per-question analysis of the grounded gain. Two regularities fix how to read Table 1. First, $P(\text{loc} \mid \text{corr})$ rises monotonically with the grounding signal—41.4 (zero-shot), 46.0 (CoT), 47.2 (correctness), 47.9 (all-trace), 50.6 (VERA), 51.3 (gold-span), 51.5 (teacher)—whereas the gap

Table 13: **Per-seed statistics for the scale study** (Table 7): Acc@GQA mean±std over three seeds for VERA and the correctness-filtered control at each student scale.

Student scale	VERA	Corr.-filt.	Δ
1B	37.2 ± 0.4	33.4 ± 0.5	$+3.8$
4B	41.0 ± 0.3	36.0 ± 0.3	$+5.0$
12B	42.4 ± 0.3	36.9 ± 0.4	$+5.5$

$G = \text{Acc@QA} - \text{Acc@GQA}$ lies in 39.8–40.4 for *every* system trained in our pipeline. Second, the advantage is concentrated on the span-annotated questions: the five controls lose 2.6–4.7 points of answer accuracy from full NExT-QA to the temporal/causal NExT-GQA subset while the VERA-family and gold-span students lose none, which also means that on the remaining (largely descriptive) questions the controls are at least as accurate: evidence-supervised training pays off specifically on questions that require evidence and buys nothing on those that do not. A per-question analysis explains the invariance of G: 31% of test questions are answered correctly but left unlocalized by *all* thirteen in-pipeline systems, the correct-but-unlocalized sets of the zero-shot student and VERA overlap at Jaccard 0.73, and these are dominated by causal *why*/*how* questions whose gold spans are short (median 3.4 s of a 44 s clip) or answerable from one frame.

Interventional grounding (Q3): implementation. Table 4 in the main text reports sensitivity, invariance, and their harmonic mean GFS under two masking implementations (mean-frame replacement and Gaussian blur, which agree within noise; the table reports the former), and the *question-swap agreement* control: on NExT-GQA videos carrying ≥ 2 questions whose gold spans are disjoint (IoU $= 0$), the fraction of question pairs for which each predicted span has higher IoU with its own question's gold span than with the other question's. A model that localizes salient events irrespective of the question scores at chance (0.5) on this control regardless of its Acc@GQA.

Acceptance-conditioned statistics and the ε_{loc} channel. Table 14 cross-tabulates verifier acceptance against answer correctness and gold-span localization on the NExT-GQA validation split (teacher rollouts at $K{=}8$; student outputs from the relabeling pass of §2.4). Because acceptance rate, conditional accuracies, and $\bar{q}$ are not free parameters—any three determine the fourth—the table prints the marginal accuracy as well, so the identity is visible: $0.46 \times 93.2 + 0.54 \times 44.7 = 67.0$ and $0.46 \times 93.2/67.0 = 0.64 = \bar{q}$. Note also that per-trace accuracy (67.0) is far below the teacher's deployed system accuracy (82.2; Table 1), because the deployed teacher runs a *single* agentic pass decoded greedily, whereas these statistics are per single rollout sampled at temperature 1.0 to expose the acceptance distribution. The converse failure mode that the verified-incorrect rate of Table 3 does not reveal—accepted, correct, yet poorly localized—amounts to 44.0% of accepted-correct teacher traces. Combining this with the 93.2% answer accuracy among accepted traces gives a pooled correct-but-ungrounded mass of $0.932 \times 0.44 \approx 0.410$ among accepted teacher traces, consistent with the approximately 0.40 input-normalized direct estimate. Roughly two-thirds of the failure mass is boundary-imprecise (true-support fraction above ρ: the V2 channel) and one-third is low-support false acceptance (the V1 channel). The same statistics computed on re-verified student outputs supply the student-side primitives of Assumption 3. True-support fractions use the gold-aligned-event definition of §A.3 on the annotated instrumentation subset.

Soundness diagnostic. Assumption 1 is stated per input. Table 15 and Figure 3(a,b) report the aggregate quantities implied by the acceptance-conditioned localization statistics. The pooled filtered-teacher grounding-gap mass is $0.932 \times 0.44 \approx 0.410$; the input-normalized direct estimate is approximately 0.40. With the per-input primitives aggregated under the same localization definition, the Proposition 2 plug-in reaches the probability ceiling, so the resulting bound is vacuous. This is a limitation of the diagnostic, not of the matched empirical comparison in Tables 1–3.

The terms of the cascade bound, tabulated. Table 16 (plotted in Figure 3c) reports every quantity entering Proposition 6 at each operating point of Table 5, the numerical value of the bound $B(\kappa)$ with $\hat{\xi}$ plugged in, the measured system grounding gap, grounded accuracy on the retained and escalated subsets separately (the test of whether the teacher dominates on the routed subset, §2.4), and the check of Assumption 4 globally and by question type. Two entries are read off the NExT-GQA block of Table 1, using that dataset's own Acc@QA so that both terms describe the same questions: the deployed teacher's global grounding gap is $G(\pi_{\text{T}}\text{-agentic}) = 0.822 - 0.423 = 0.399$, and the student's is $0.810 - 0.410 = 0.400$. Neither may be computed by subtracting Acc@GQA from the NExT-QA column, which reports a different question population; both are dominated by correct-but-unlocalized answers that no system on this benchmark avoids. Three things the table

Table 14: **Acceptance-conditioned statistics** on the NExT-GQA validation split. Left column: teacher traces; right column: student outputs under offline re-verification (the measurement behind Assumption 3). Localization uses the official $IoP \geq 0.5$ rule; median IoP and IoU are also reported. Per-trace answer accuracy is measured from single rollouts sampled at temperature 1.0 under the learned controller, rather than from the deployed teacher's greedy single agentic pass reported in Table 1. Student outputs are decoded greedily on the validation split, which is why their marginal accuracy (76.8) is below the test-split figure of Table 1 (81.0).

	Teacher traces	Student outputs
Acceptance rate a (per trace/output)	0.46	0.52
Answer accuracy (marginal, per trace/output)	67.0	76.8
Answer accuracy | accepted	93.2	90.1
Answer accuracy | rejected	44.7	62.4
Verified yield among correct traces ($\bar{q}$)	0.64	0.61
P($IoP \geq 0.5$ | accepted $\wedge$ correct)	56.0	55.0
P($IoP \geq 0.5$ | rejected $\wedge$ correct)	41.2	43.5
Median IoP | accepted $\wedge$ correct	0.71	0.68
Median IoU | accepted $\wedge$ correct	0.61	0.58
P($\neg$localized | accepted $\wedge$ correct), IoP rule	44.0	45.0
of which boundary-imprecise (V2 channel)	29.3	30.0
of which low-support false-accept (V1 channel)	14.7	15.0

Table 15: **Soundness diagnostic** on the NExT-GQA instrumentation subset. The direct pooled quantity is determined by Table 14. The input-normalized value is the corresponding per-input aggregation. The Proposition 2 plug-in is clipped to 1 by definition and is therefore vacuous on this benchmark.

Quantity	Value
P($IoP \geq 0.5$ | accepted $\wedge$ correct), pooled	0.56
P($\neg$localized | accepted $\wedge$ correct), pooled	0.44
P(correct | accepted), pooled	0.932
Direct pooled $G(\pi_{\mathrm{T}}^{\mathrm{Ver}}) = P(\mathrm{correct} \wedge \neg\mathrm{loc} \mid \mathrm{accepted})$	0.410
Direct $\hat{G}(\pi_{\mathrm{T}}^{\mathrm{Ver}})$, input-normalized	≈ 0.40
Proposition 2 plug-in bound after clipping	1.00 (vacuous)

makes visible. First, the student-side soundness plug-in clips at $\delta_{\mathrm{s}} = 1$, so the cascade bound is vacuous through 50% escalation and remains very loose even at 80%, against a measured system gap of ≈ 0.40 (Remark 4); its limiting value as $\kappa \to 1$ is $\hat{\xi} + G_{\mathrm{T}}^{\mathrm{esc}}$, so on a benchmark where the escalation target itself has a gap of 0.40 no bound of this form can certify a small absolute gap. Second, the measured gap is essentially flat in κ: the cascade raises Acc@GQA and Acc *together* (student and teacher have nearly identical gaps), so it improves grounded accuracy without shrinking the correct-but-ungrounded mass—the intended reading of Table 5 is the former, not the latter. Third, Assumption 4 is satisfied only trivially because $\delta_{\mathrm{s}} = 1$; it therefore carries no useful empirical information here. The teacher nevertheless dominates the student on the escalated subsets by 1.6–5.0 Acc@GQA points (most at the cost-driven 10–20% operating points), which is what makes escalation worthwhile empirically.

What does $\hat{v}_\psi$ predict? The ECE of 0.045 (Figure 2c) is computed against the re-run verifier's decision Y, so a well-calibrated $\hat{v}_\psi$ could in principle predict *acceptance* well while predicting *groundedness* poorly—Table 14 shows 44–45% of accepted-and-correct outputs are unlocalized. Table 17 therefore evaluates $\hat{v}_\psi$ on the annotated instrumentation subset against three targets: (i) the verifier decision Y, (ii) the joint event $\mathbf{1}[\hat{y}{=}y^\star \wedge \mathrm{loc}(\hat{g}, g^\star){=}1]$, and (iii) identification of correct-but-unlocalized outputs among correct ones. The comparison quantifies how much of the routing gain of Table 5 comes from catching genuinely ungrounded outputs versus outputs the particular verifier would reject.

Dose–response test of Proposition 8. Table 18 sweeps the alignment weight λ_r with stages (i)/(ii) of §2.4 held fixed and reports the achieved $\mathcal{L}_{\mathrm{align}}$, its image $\Phi(\mathcal{L}_{\mathrm{align}})$ under Lemma 2, and the measured $\hat{\xi}$ with bootstrap intervals. The proposition bounds ξ from above by a quantity that falls with λ_r, which is consistent with—but does not imply—a fall in the measured $\hat{\xi}$; we therefore report the sweep as an observation and draw no inference about a floor from it (§A.10).

Table 16: **Terms of Proposition 6** at each operating point on NExT-GQA (test split; $\hat{v}_\psi$-routing). $\hat{\xi} = \mathbb{E}|\hat{v}_\psi - Y|$ is the measurable upper bound on the ℓ_1 calibration error (a property of the head, hence constant across columns); δ_s is the student-side bound of Proposition 2 on the retained set; it clips at 1 at every reported operating point; G_T^{esc} is the teacher's gap on the escalated subset; B_k is the bound evaluated at the class's right endpoint κ_k (Definition 2), which is where it is tightest and where it is well defined—B is *not* constant across a routing-equivalence class, since it depends on κ explicitly through $r_\kappa(1-\kappa)$. Acc@GQA is reported on retained (student) and escalated (teacher vs. student on the same inputs) subsets. As a consistency check, the retain-weighted mixture of the retained and escalated *student* columns reproduces the student's 41.0 in every column, and the mixture of retained student with escalated teacher reproduces the system row.

Escalation rate $1 - r_\kappa$	10%	20%	50%	80%
κ (class representative)	0.16	0.26	0.51	0.82
r_κ	0.90	0.80	0.50	0.20
ECE / $\hat{\xi}$	0.045 / 0.24	0.045 / 0.24	0.045 / 0.24	0.045 / 0.24
δ_s (bound on retained set)	1.00	1.00	1.00	1.00
$G_T^{esc}(\kappa)$ (teacher gap, escalated subset)	0.43	0.42	0.41	0.40
Retained-rejection term $r_\kappa(1-\kappa_k)$	0.76	0.59	0.25	0.04
B_k (Proposition 6, class endpoint)	1.94 (vacuous)	1.72 (vacuous)	1.19 (vacuous)	0.80 (loose)
Measured $G(\text{cascade})$	0.401	0.401	0.401	0.401
Acc@GQA, retained subset (student)	41.8	42.6	44.5	48.0
Acc@GQA, escalated subset: teacher / student	38.8 / 33.8	38.6 / 34.6	39.7 / 37.5	40.9 / 39.3
System Acc@GQA (deployed cascade)	41.5	41.8	42.1	42.3
Assumption 4 holds (global / worst question type)	yes / yes (trivial)	yes / yes (trivial)	yes / yes (trivial)	yes / yes (trivial)

Table 17: **Calibration and discrimination of $\hat{v}_\psi$ by target** on the NExT-GQA instrumentation subset (stage-(ii) relabeled head; student outputs). AUROC is threshold-free ranking; ECE uses 15 equal-mass bins. Row (iii) is restricted to outputs with $\hat{y} = y^\star$ and asks whether low $\hat{v}_\psi$ identifies the correct-but-unlocalized ones.

Target	AUROC	ECE
(i) verifier acceptance Y	0.87	0.045
(ii) joint correct $\wedge$ localized	0.72	0.12
(iii) $\neg$localized among correct ($1 - \hat{v}_\psi$ as score)	0.60	—

Table 18: **Alignment weight vs. calibration** on NExT-GQA (stage (ii) relabeled head; 95% bootstrap CIs). $M = 255$ negatives per anchor, temperature $\iota = 0.2$; Φ from Lemma 2; $a(\theta)$ is the directly measured mean distance to the positive belief encoding, which Φ upper-bounds. Note that Φ is loose at small ι (the maximum distance on the unit sphere is 2); an upper bound that falls with λ_r does not imply that the measured $\hat{\xi}$ falls, so the sweep below is descriptive and is not a test of Proposition 8 (§A.10).

λ_r	0	0.1	0.3	1.0 (default)	3.0
$\mathcal{L}_{\text{align}}$ (final)	4.6	2.1	1.3	0.9	0.8
$\Phi(\mathcal{L}_{\text{align}})$ (Lemma 2)	1.90	1.62	1.52	1.46	1.45
$a(\theta)$ (measured)	1.21	0.93	0.78	0.71	0.70
ECE	0.061	0.053	0.048	0.045	0.044
$\hat{\xi}$ [95% CI]	0.29 [0.28, 0.30]	0.26 [0.25, 0.27]	0.25 [0.24, 0.26]	0.24 [0.23, 0.25]	0.24 [0.23, 0.25]
Acc@GQA	40.4	40.7	40.9	41.0	40.8

Probe protocol for Table 6. Probes are ℓ_2-regularized logistic-regression classifiers on frozen $z_{\theta,k}$, trained, validated, and tested on video-disjoint splits of the 1,000-clip instrumentation subset: no probe-test video or question appears in probe training, and probe-test *videos* are additionally disjoint from the distillation corpus. Event-order and postcondition labels come from two independent annotations per clip (pool of three annotators; agreement Cohen's $\kappa = 0.78$). Probe capacity and regularization are selected on the probe-validation split only. Table 19 adds bootstrap CIs for the headline rows and a *shuffled teacher-belief targets* control, in which each anchor's positive $e(b_{t(k)})$ is replaced by a belief encoding drawn from a different input at training time: alignment trained against shuffled targets should not transfer belief structure, so probe accuracy near the no-alignment row is the expected outcome if Table 6's gains are genuine.

Ablations reported in Table 8. (i) The learned controller vs. the teacher's fixed thresholds, with distilled Acc@GQA plotted against measured (u, N_{cov}) to test the coverage channel of Proposition 4

Table 19: **Probe CIs and shuffled-target control** for Table 6 (95% bootstrap over probe-test videos).

Representations from	Event order↑	Postcond.↑
VERA without $\mathcal{L}_{\text{align}}$	64.8 [61.8, 67.7]	48.9 [45.8, 52.0]
VERA (full)	74.5 [71.8, 77.2]	60.7 [57.7, 63.7]
trained with shuffled belief targets	65.5	49.6

(Table 8 and Figure 3(f)); (ii) loss components: answer-only target, +evidence, and λ_r=0; (iii) inverse-yield debiasing on/off (the input-level weight $\hat{w}$ set to 1; within-input tilt weights ω retained); (iv) the tilt γ; (v) the rollout budget K; (vi) the InfoNCE negative sources of Eq. (7); (vii) cross-fitted vs. same-sample propensities, the measured cost of the bias analyzed in Remark 6; (viii) label-based vs. verifier-only selection of $(\eta, \gamma, \kappa, \lambda_\cdot)$; and (ix) the mean-yield vs. coverage-sensitive controller objective. The self-verifier's training distribution (stage (i) only vs. +stage (ii)) is reported as the calibration shift of Figure 2(c). We also report the aggregate grounding diagnostic in Table 15. The acceptance-conditioned localization rates imply a direct filtered-teacher gap of approximately 0.40; the Proposition 2 plug-in saturates at 1 and is therefore not used as a non-vacuous empirical certificate.

D Robustness to Perception Noise

We stress-test the perception front-end via caption dropout (rate ρ_c) and detector confidence jitter (scale s), reporting grounded accuracy for VERA, CoT distillation, and the teacher (Table 20). Because the clean-condition soundness plug-in is already vacuous, the perturbation study is interpreted empirically rather than as a bound-tracking experiment. Both hold: under caption dropout VERA loses 4.6 Acc@GQA against CoT distillation's 7.1 (teacher: 3.4), and under detector noise 2.8 against 4.4 (teacher: 2.2).

Verifier-side error analysis. Assumption 1(V1) remains a conditional, per-input statement, so a pooled false-accept rate alone cannot establish a non-vacuous grounding guarantee. The direct accepted-trace localization failure is substantial, and the Proposition 2 plug-in saturates. We therefore retain the proposition as a population-level diagnostic and use the perturbation study below only to test empirical robustness of VERA, rather than claiming that the measured verifier primitives certify temporal localization.

Table 20: **Robustness to perception noise** on NExT-GQA. We report empirical Acc@GQA only; the Proposition 2 plug-in is already vacuous in the clean condition, so no bound values are presented for the perturbations.

	clean	caption dropout ρ_c=0.3	detector noise s=0.2
CoT distillation	34.2	27.1	29.8
VERA (ours)	41.0	36.4	38.2
Teacher (agentic)	42.3	38.9	40.1

E Qualitative Case Study

A worked example (intuition). For q ="*Why does the boy run to the table?*", a trace hypothesizing "*because a cake is brought out*" *verifies*—a "cake placed on table" event is detected and precedes the running—and carries evidence spans on those frames; a competing trace answering "*to eat cake*" from prior expectation, without localizing any trigger, *fails* verification. The student learns from the first to answer *and* to point at the trigger; the second, rather than being wasted, becomes a negative example teaching the self-verifier what an unsupported answer looks like. Distilling the unverified trace would teach exactly the correct-but-ungrounded behavior we want to avoid—the concrete sense in which verification, not the chain-of-thought format, supplies the grounding.

F A Second Instantiation, Sketched: Formal Mathematical Reasoning

To make the domain-independence of the recipe's *interfaces* concrete without extending the empirical claims beyond VideoQA, we describe a corresponding instantiation for stepwise mathematical reasoning with a symbolic checker. The mapping follows the interfaces of §2. The *teacher* is an LLM agent producing stepwise derivations; the *belief state* b_t is the current subgoal and known facts; *actions* are tactic or lemma selections; and *observations* are the checker's state after applying

each step. The *verifier* is the symbolic checker itself, with the per-step postcondition requiring that each step be a valid inference from its cited premises. For a sound formal checker, this substantially reduces verifier error and correspondingly tightens the soundness term in Proposition 2. *Typed evidence* consists of the premise citations associated with each step, making verification re-runnable on student outputs as in §2.4 and supporting the same stage-(ii) relabeling and verify-then-escalate mechanism. The analogue of the grounding gap is a final-answer-correct derivation containing invalid or unsupported intermediate steps. Under these interface definitions, the deliberate–certify–amortize stages of Algorithm 1 transfer directly. This appendix establishes structural compatibility of the framework's interfaces across domains; all empirical claims in this paper remain restricted to the evaluated VideoQA settings.

G Extended Related Work

Grounded VideoQA. NExT-QA (Xiao et al., 2021), IntentQA (Li et al., 2023), and EgoSchema (Mangalam et al., 2023) probe temporal and causal understanding; NExT-GQA (Xiao et al., 2024), ReXTime (Chen et al., 2024), and CG-Bench (Chen et al., 2025a) measure whether answers are localized to correct evidence through a joint answer-plus-span criterion, and Ego4D-NLQ (Grauman et al., 2022) measures localization alone. Agentic video reasoners with explicit verifier roles (Liu et al., 2026b) supply exactly the interface our teacher-side pipeline consumes. Video LLMs (Maaz et al., 2024; Ren et al., 2024; Team et al., 2024; 2025) push accuracy, and grounded systems—SeViLA (Yu et al., 2023), MoReVQA (Min et al., 2024), MUPA (Dang et al., 2025), LeAdQA (Dong et al., 2026)—attack localization directly; these supply our external baselines (§3). Two recent lines bear directly on *what* a predicted span should be evidence *of*. (Seo et al., 2026) show that much of the grounding gap on NExT-GQA traces to grounders that fix video features before they receive question semantics, and condition localization on question intent so that the predicted span is question-discriminative rather than merely salient; VERA's typed evidence is question-conditioned by construction (each span is cited by a belief state that names its role in answering q), but we do not compare against their model directly, and our question-swap control (Table 4) is the evaluation-side response to their diagnosis. (Liu et al., 2025) decompose a question into an entailment tree of hypotheses and verify each against video fragments before answering; this is the closest published analogue of the teacher's per-step postcondition verification, with the difference that their verification is executed at inference, whereas VERA uses it as a training-time curation signal and amortizes it into the deployed student—a direct cost–grounding comparison against entailment-tree reasoning is a natural experiment we have not run. Agentic role decomposition is the closest published analogue of our teacher-side workflow: VideoMind (Liu et al., 2026b) chains planner, grounder, verifier, and answerer roles through a chain of LoRA adapters for long-video reasoning—like our agentic teacher, it buys grounding with inference-time deliberation, which is exactly the cost–grounding trade-off VERA amortizes rather than executes per query; Table 1 places its single-agent operating point on the same cost–grounding axes as our teacher/student decomposition; a full frontier comparison is future work. (Luo et al., 2026) diagnose *visual thinking drift*—chains of thought that detach from the frames—and counter it with an evidential-grounding reward during reinforcement learning; VERA attacks the same failure mode with an *external* trace verifier used for curation, distillation, and deployment-time self-verification rather than for reward shaping, and the two signals are complementary. VITED (Lu et al., 2025) is the closest single-pass counterpart of our student: it trains a temporally aware VLM to generate an answer together with a chain of temporal evidence spans, using evidence chains synthesized by searching for intervals that maximize the likelihood of the gold answer. Its curation signal is therefore answer-driven, i.e. the correctness-filtered boundary of Proposition 1, whereas VERA curates by an external verifier and, because its evidence is typed, can re-run that verifier on student outputs; Table 1 reports a direct comparison on NExT-GQA under the official evaluator. ViQAgent (Montes & Lozano, 2026) is a further verifier-equipped VideoQA agent, cross-checking grounding timeframes with an open-vocabulary detector at inference; it is a candidate teacher for the interface of §2, alongside VideoMind. NeuS-QA (Shah et al., 2026) takes verification furthest toward formal methods: it translates the question into a temporal-logic specification, builds a video automaton over frame-level propositions, and model-checks segments *before* answering—verification at retrieval time, training-free. Relative to it, VERA uses per-step postcondition verification as a training-time curation signal and amortizes the verifier into the deployed model; a temporal-logic checker is itself a candidate v_t for our interface, with a near-formal ε_v in Assumption 1. Traceable-evidence evaluation and training (Wang et al., 2026a) couples reasoning with joint localization supervision and traceable evidence; our typed-evidence supervision and in-

terventional GFS protocol are the distillation-side counterparts, and its evaluation methodology is a natural additional testbed for VERA's evidence-traceability claims. A closely related recent line makes *explicit* evidence part of the reasoning output itself: Open-o3 Video trains grounded video reasoning with explicit spatio-temporal evidence (Meng et al., 2025), Evidence-Backed VideoQA requires answers together with precise spatio-temporal evidence including segmentation masks (Wang et al., 2026b), Video-in-the-Loop interleaves span-grounded reasoning over long videos (Wang et al., 2025), MultiHop-EgoQA requires multi-hop evidence aggregation in long egocentric video (Chen et al., 2025b), and intra-video temporal-aware retrieval grounds answers in retrieved timestamped evidence (Shafiq et al., 2026). On the benchmark side, EG-VQA targets *verifiable* VideoQA with annotated supporting temporal evidence (Huang et al., 2026) and CaST-Bench evaluates causal-chain-grounded spatio-temporal reasoning (Zhang et al., 2026); both are natural additional testbeds for VERA's typed-evidence and multi-step-belief claims, and we view evaluation on them as important future validation beyond single-span localization. These systems and benchmarks make evidence explicit at inference or in annotation; VERA is complementary and differs in *certifying* evidence with an external verifier during training, distilling the certified behavior into a single pass, and deploying a calibrated estimate of whether the evidence would survive verification. The persistent grounding gap motivates making grounding part of the reasoning process itself.

Amortizing deliberation. Compressing multi-step "System-2" computation into a single pass has been studied for text reasoning, most directly by System-2 distillation (Yu et al., 2024), which filters self-generated traces by output consistency before distilling. A concurrent line filters distillation data by *verification*: self-verified distillation uses the model's own verification as a synthetic-data filter (Lee & Liang, 2026), execute–distill–verify paradigms use multi-agent consensus to vet experiences before committing them (Zhu et al., 2026), and, closest to our setting in *what* is transferred, DocVAL (Mohammadshirazi et al., 2025) distills validated chains of thought from a large teacher into a compact student for grounded document VQA, using a multi-module validator that enforces answer correctness and geometric consistency of the cited region. DocVAL's validator conditions on answer correctness, so at the filtering-operator level it sits between cases 2 and 3 of Proposition 1 (our $\mathrm{Ver}\wedge$correct corpus); VERA differs in using a verifier that is answer-label-free, in distilling belief-state representations alongside outputs, and in amortizing the validator itself into the student. VISD (Lin et al., 2026) is the nearest self-distillation counterpart on video: a video-aware judge decomposes reasoning quality into answer correctness, logical consistency, and spatio-temporal grounding and modulates token-level supervision inside an RL loop. Its grounding dimension conditions on answer correctness and shapes gradients rather than curating a corpus, so at the operator level it is a judge-weighted relative of the $\mathrm{Ver}\wedge$correct boundary of Proposition 1; VERA's selection signal is an external answer-label-free verifier, applied as trace-level curation whose acceptance is then itself amortized. On the verification side, (Kumar, 2026) study post-hoc self-verification of timestamped citations in Video-LLM answers and ablate the verifier's components; their finding that verifier design, not merely its presence, determines which hallucinated citations are caught is the empirical counterpart of Corollary 2, and their ablation protocol is a natural complement to our perception-noise study (Appendix D). Selective invocation of expensive reasoning has also been pursued *inside* a single model: VideoAuto-R1 (Liu et al., 2026a) learns to think once and answer twice, deciding per query whether deliberation is warranted; VERA's cascade (§2.4, Figure 2, Table 5) externalizes that decision into a calibrated estimate of *verifier* acceptance with an end-to-end bound (Proposition 6) and a head-free verify-then-escalate control, and the two mechanisms could compose. VidBridge-R1 (Chen et al., 2026) improves RL-based video understanding through intermediate proxy tasks bridging QA and captioning; relative to such intermediate-supervision signals, the controller of Eq. (2) consumes the verifier's certificate directly, and whether proxy-task shaping and verifier reward are complementary is an open question. VERA differs in the curation signal (an *external, perception-grounded* verifier over video events rather than self-consistency or self-verification), in the supervision (typed evidence spans and representation-level belief alignment, not only answers), in amortizing the verifier into the deployed model, and in providing a conditional grounding bound (Theorem 1 shows consistency- or correctness-based curation admits no analogue).

Verifier-guided self-training and control. STaR and rejection-sampling fine-tuning bootstrap from *correct* samples (Zelikman et al., 2022; Yuan et al., 2023); V-STaR additionally trains a verifier with DPO (Hosseini et al., 2024); process- and outcome-level verifiers score reasoning steps (Cobbe et al., 2021; Lightman et al., 2024). Our controller-then-distill loop inherits the structure of expert iteration (Anthony et al., 2017) and policy distillation (Rusu et al., 2015; Ross et al., 2011), with the verifier—not answer correctness—as the improvement operator; the correctness-filtered baselines in §3 isolate

exactly this difference. Self-refinement and reflection (Madaan et al., 2023; Shinn et al., 2023; Weng et al., 2023) improve a single trajectory; we use verification as a *supervision and curation* signal for a policy, a student, and a self-verifier.

Reasoning distillation. Knowledge distillation (Hinton et al., 2015) and its reasoning-specific forms (Hsieh et al., 2023; Ho et al., 2023; Magister et al., 2023) transfer chains-of-thought. In VideoQA specifically, VideoDistill (Zou et al., 2024) distills language-aware visual semantics so that a student attends to question-relevant frames; it transfers *what to look at* but produces no evidence output, so it cannot be scored on a joint answer-plus-span metric, and it is complementary to trace curation rather than an alternative to it. Our target is a verification-tilted, evidence-bearing trace distribution, and we additionally align internal representations to belief states—closer in spirit to feature-level distillation than to text imitation.

Selective prediction, cascades, and faithfulness evaluation. The self-verification head connects to selective classification (Geifman & El-Yaniv, 2017), calibration (Guo et al., 2017), and language models' calibrated self-evaluation (Kadavath et al., 2022); within grounded VideoQA specifically, (Xu et al., 2024) calibrate answer confidence with question guidance, whereas $\hat{v}_\psi$ predicts *verifier* acceptance of typed evidence rather than answer confidence—exactly the distinction our max-probability routing control isolates. The escalation cascade connects to cost-aware LLM cascades and routing (Chen et al., 2023). Our GFS metric is the interventional, video-native extension of comprehensiveness/sufficiency rationale metrics (DeYoung et al., 2020; Jacovi & Goldberg, 2020), hardened with randomized-ablation controls. Preference optimization for the controller follows (Rafailov et al., 2023). Conceptually, a fast, self-verifying understanding module is a natural companion to world-model agendas (Ha & Schmidhuber, 2018; LeCun et al., 2022).

Relative to all of the above, VERA differs on three axes: the curation signal is an *external, perception-grounded* verifier rather than answer correctness, self-consistency, or self-verification; what is transferred includes typed evidence and span-indexed belief representations, not only answers; and the verifier is amortized *into* the deployed model rather than used only for training-time curation or test-time reranking—with Theorem 1 showing that correctness- or consistency-based curation admits no analogous grounding guarantee.