FROM DECOMPOSITION TO INTEGRATION: STAGE-SPECIALIZED ON-POLICY DISTILLATION FOR VIDEO DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained video diagnosis requires identifying the relevant subject, discovering subtle differences, and deciding which differences constitute a defect. We present a framework that turns these interdependent capabilities into structured supervision and integrates them within one model. A multipass annotation pipeline decomposes reference-conditioned diagnosis into Subject Binding, Multidimensional Comparison, and Issue Filtering, producing compact traces that connect visual evidence to diagnostic decisions. We then introduce Stage-Specialized On-Policy Distillation (Stage OPD): the student generates a complete trajectory, and each specialist provides feedback for its stage under the student's own preceding outputs. Annotated-trace replay supplies complementary supervision for content, stage transitions, and diagnostic decisions. With 2,160 training traces, the resulting 9B student improves both precision and recall on a human-reviewed benchmark, raising issue-detection F_1 from 36.68% to 41.88%. The student exceeds a three-call SFT chain in F_1 and approaches the 27B reference within 0.31 F_1 points while retaining a single 9B model at inference. Overall, stage OPD organizes specialized supervision through decomposition and integrates the resulting diagnostic capabilities into one model.

1 INTRODUCTION

A generated video can look convincing while getting a small but consequential detail wrong: a player's shirt number changes, a product loses its logo, or a character's hairstyle differs from the supplied reference. Diagnosis requires identifying the subject and its reference, inspecting attributes across frames, and deciding whether a change conflicts with the generation prompt. The evidence is often sparse: a discrepancy may occupy a small region in only two frames, surrounded by otherwise consistent content. Multiple subjects and changing viewpoints make this evidence-to-decision link difficult.

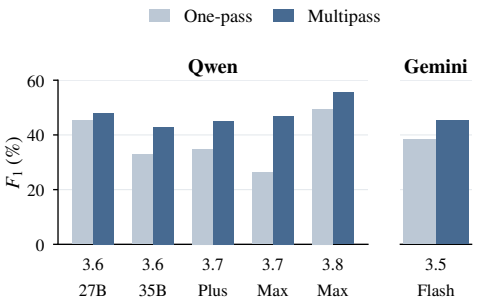

Figure 1: One-pass and multipass video diagnosis across six Qwen and Gemini models. Each model uses direct diagnosis or the same three-stage diagnostic pipeline.

Explicit decomposition offers a practical way to organize these demands. Figure 1 contrasts direct diagnosis with an explicitly decomposed procedure across several model families. We instead ask three focused questions: *who should be compared, what differs, and which differences should be reported?* Our multipass annotation pipeline first establishes subject–reference bindings, then compares complementary attribute dimensions in parallel, and finally filters and consolidates the observations. Shared bindings preserve subject correspondence; focused comparisons expose small differences; filtering excludes permitted changes and occlusions. Each stage exposes intermediate decisions that provide evidence-linked process supervision.

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.

The effectiveness of a multipass pipeline raises a learning question: *how can its specialized capabilities be integrated into a single model?* External decomposition requires repeated model calls and intermediate-state exchange. A unified model should perform the complete diagnosis in one call. To establish a shared representation, we convert the pipeline outputs into 2,160 compact diagnostic traces. The traces preserve bindings, paired evidence, and observation-to-event links, including unchanged attributes and permitted differences that should not be reported. Supervised fine-tuning (SFT) learns this representation and the dependencies among its stages. Further improvement requires feedback on student-generated intermediate histories.

We propose *Stage-Specialized On-Policy Distillation* (Stage OPD) to provide this feedback. On-policy distillation trains on student-generated responses (Agarwal et al., 2024; Gu et al., 2024); Stage OPD additionally organizes the source of supervision by diagnostic stage. Three specialists are trained from the compact traces, emphasizing Subject Binding, Multidimensional Comparison, and Issue Filtering. The student generates a complete trajectory, and each frozen specialist scores the tokens in its assigned stage under the same student-generated history. This preserves one shared reasoning context while matching each stage to focused expertise. Annotated-trace replay supplies complementary supervision for descriptive content, stage transitions, and the sparse decisions governing the final report. The central idea is to use decomposition to specialize supervision during training and integrate the resulting capabilities within one student at inference.

We evaluate this approach on a human-reviewed reference-to-video benchmark with issue-level matching. The resulting 9B student improves precision from 51.22% to 56.32% and recall from 28.57% to 33.33%, raising F_1 from 36.68% to 41.88%. It approaches the 27B reference within 0.31 F_1 points and exceeds the three-call SFT chain's 40.47% while producing the entire diagnosis in one call. Compared with prompting SFT to execute the stages separately, the learned student improves reporting performance. Our analyses examine specialist supervision, diagnostic decisions, and the benefit of integrating a three-call process into one generation. Our contributions are:

- **Stage-specialized on-policy distillation.** We introduce a framework in which specialists supervise their corresponding spans of a complete student-generated trajectory. Structure- and decision-aware replay complements this transfer, yielding one model for end-to-end video diagnosis.
- **Multipass annotation and compact process supervision.** We develop a reusable pipeline with shared subject bindings, parallel multidimensional comparisons, and explicit issue filtering, and construct 2,160 traces that preserve the evidence and decisions connecting these stages.
- **A reviewed benchmark and empirical analysis.** We establish issue-level evaluation of reference-conditioned video diagnosis and demonstrate gains in both precision and recall, supported by component studies, specialist analysis, and training dynamics.

2 Related Work

Knowledge and On-Policy Distillation. Distillation transfers teacher behavior to a student (Hinton et al., 2015); sequence-level distillation extends this idea to autoregressive outputs (Kim & Rush, 2016). DAgger and scheduled sampling address training–inference distribution shift (Ross et al., 2011; Bengio et al., 2015). GKD learns from student-generated mistakes (Agarwal et al., 2024), and MiniLLM develops reverse-KL distillation for language models (Gu et al., 2024). Video-OPD applies dense teacher feedback to video understanding (Li et al., 2026), while MOPD routes prompts to specialist teachers (Ma et al., 2026). Stage OPD specializes teachers within one student trajectory.

Decomposition and Process Supervision. Chain-of-thought prompting exposes intermediate reasoning (Wei et al., 2022; Zhang et al., 2024); least-to-most and decomposed prompting organize a complex task into simpler calls (Zhou et al., 2023; Khot et al., 2023). Distilling Step-by-Step uses teacher rationales as supervision for smaller models (Hsieh et al., 2023), and process supervision evaluates intermediate reasoning steps (Lightman et al., 2024). We construct diagnostic traces through decomposition, then distill their stages into one model.

Video Understanding. MVBench and Video-MME evaluate broad video-language capabilities (Li et al., 2024; Fu et al., 2025), while LongVideoBench emphasizes reasoning over long interleaved contexts (Wu et al., 2024). Further video benchmarks examine fine-grained distinctions in video

understanding (Liu et al., 2024; Cai et al., 2024). Our diagnosis task additionally requires subject–reference comparison and subject-specific defect reporting.

Video Generation Assessment. VBench, VBench++, and VBench-2.0 measure complementary aspects of generated video quality and faithfulness (Huang et al., 2024; 2025; Zheng et al., 2025); T2V-CompBench targets compositional consistency (Sun et al., 2025). VideoScore and VideoScore2 learn fine-grained assessment from human feedback (He et al., 2024; 2025), complementing human-aligned benchmarking and reward learning (Han et al., 2025; Liu et al., 2025a). Phantom, Phantom-Data, and OpenS2V-Nexus emphasize subject consistency in generation and data construction (Liu et al., 2025b; Chen et al., 2025; Yuan et al., 2025). Our diagnostic reports identify the affected subject and attribute, making false alarms and missed defects measurable.

3 Task, Data, and Benchmark

3.1 Reference-Conditioned Video Diagnosis

An input $x = (p, \mathcal{R}, V)$ contains a generation prompt p, reference images $\mathcal{R}$, and sampled video frames V. The output is a set of subject-consistency issues $\mathcal{Y}$, each specifying a defect. The correct output can be empty. A reportable issue is an unintended change to a referenced subject's identity or appearance; requested transformations, pose changes, occlusion, and camera motion are not themselves defects. The task requires identifying *what* changed relative to *which* reference, then deciding whether that difference should be reported. Each diagnosis therefore connects the affected subject, comparison evidence, and defect description.

We organize the diagnosis into Subject Binding z^B, Multidimensional Comparison z^C, and Issue Filtering z^F. The final stage selects and consolidates observations into reportable events. Writing $z = (z^B, z^C, z^F)$, the trace supports the factorization

$$p_\theta(z, \mathcal{Y} \mid x) = p_\theta(z^B \mid x)\, p_\theta(z^C \mid x, z^B)\, p_\theta(z^F \mid x, z^B, z^C)\, p_\theta(\mathcal{Y} \mid x, z). \tag{1}$$

The same representation connects annotation, specialist training, and student inference.

3.2 Multipass Annotation

Our pipeline uses Gemini-3.5-Flash and Qwen3.8-Max to annotate archived reference-to-video examples through three stages: *Subject Binding*, *Multidimensional Comparison*, and *Issue Filtering*. All stages share the same visual inputs and sampled frame sequence, allowing each observation to be traced to its subject and visual evidence. Figure 2 shows the information flow with an example from training.

Subject Binding. The first stage associates video subjects with reference identifiers and records their type, primary or secondary status, and distinguishing appearance. Shared bindings maintain subject correspondence even when another character dominates the clip, distinguishing absence from appearance changes.

Multidimensional Comparison. For each primary subject, the pipeline creates separate comparison tasks for shape and physical structure, color and material, and either appearance or product markings. Secondary subjects receive a grouped comparison. After binding, these dimension-specific calls run independently in parallel and their outputs are joined before filtering. Each call records paired reference and video evidence, subject attributes, observed differences, and prompt-based expectedness. Empty comparisons preserve evidence of consistency. Figure 2 shows all three dimensions and the join before filtering.

Issue Filtering. Each comparison labels whether the generation prompt permits an observed difference. The filtering stage removes prompt-permitted changes, selects reportable subject differences, and preserves excluded observations as ignored evidence for the compact trace. In Figure 2, Chloe's unrequested hairstyle change is retained as a subject-consistency issue, while Ava's clothing change is excluded because the prompt specifies black nightwear. Compact-trace construction subsequently groups descriptions of the same subject-level event while preserving their source links.

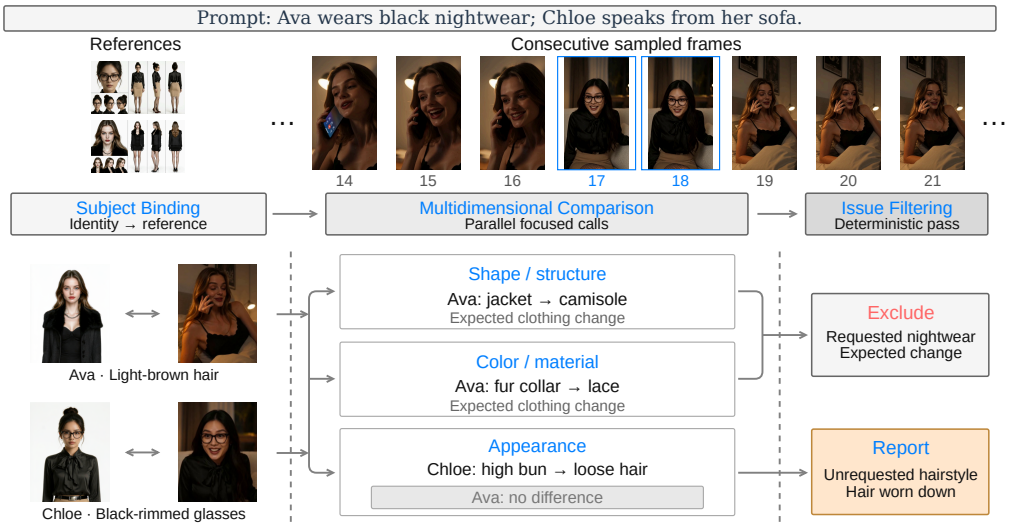

Figure 2: Multipass annotation through Subject Binding, Multidimensional Comparison, and Issue Filtering. Bound subjects are inspected across three parallel dimensions. Ava's requested clothing changes are excluded, while Chloe's unrequested hairstyle change is reported; the gray branch records no appearance difference for Ava. The reported defect concerns the mismatched hairstyle.

Visual annotation requires one binding call plus focused comparison calls; filtering uses the recorded comparisons and rewriting uses text only. This differs from the learned three-call chains in Section 5. Appendix F provides prompts and an abridged training example.

3.3 Compact Process Supervision

We convert the multipass annotations into one compact autoregressive training target through constrained text-only rewriting. The rewriter condenses identity descriptions and paired evidence while preserving subject assignments, visual facts, expectedness, and source-to-event links. Each reportable source change belongs to one event, preserving its subject, evidence, and source links.

Table 1: Training and reviewed benchmark statistics.

Process supervision	
Compact traces	2,160
Comparison blocks	15,117
No-difference blocks	9,137
Change records	8,485
Reviewed benchmark	
Video cases / issues	50 / 49
Positive / negative cases	41 / 9

After filtering Gemini- and Qwen-derived annotations, we retain 2,160 training traces (Table 1), excluding all known evaluation identifiers. The traces contain 15,117 comparison blocks and 8,485 change records, including expected changes. Of the comparison blocks, 9,137 contain no differences. Thus, a video with a reportable issue still teaches the model which other attributes remain consistent. Appendix A gives filtering details and data coverage statistics.

3.4 Benchmark and Metrics

We evaluate on 50 videos with 49 human-reviewed issues, covering reference-conditioned subject consistency. Forty-one cases contain an issue and nine contain none. Review checks the affected subject and defect semantics against the original media. Inputs retain the generation prompt, references, and saved frame sequence.

A fixed text-only judge determines whether each predicted issue and reviewed issue concern the same subject and defect. A fixed one-to-one assignment matches compatible predictions and reviewed issues. Unmatched predictions, including duplicate reports, are false positives; unmatched reviewed issues are misses. We report precision, recall, and F_1 from pooled counts over three decoding seeds. Together, these metrics measure the reliability of predicted reports and coverage of reviewed defects. Appendix C details the matching protocol.

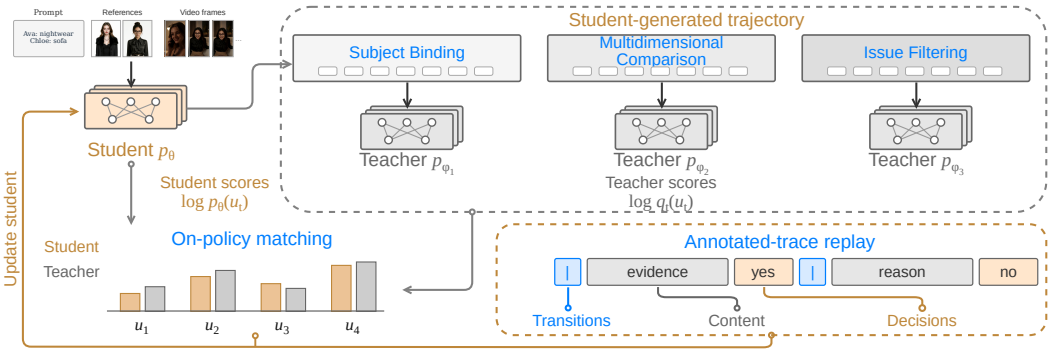

Figure 3: Stage-specialized on-policy distillation. Frozen specialists score their assigned stages under the student's history $h_t = (x, u_{<t})$. Paired probability bars illustrate on-policy feedback; the replay strip distinguishes content, delimiters, and decisions. Both signals update the student, which performs all stages at inference.

4 Stage-Specialized On-Policy Distillation

Figure 3 illustrates how Stage OPD assigns specialist feedback along a complete student-generated trajectory. One student maintains the shared history across stages and produces the full diagnosis at inference.

4.1 Specialist Initialization

Student SFT uses complete compact traces. Three specialists share its base model and training data but supervise assigned token spans. For annotation $u^\star$ and specialist token set $\mathcal{I}_k^\star$, the objective is

$$\mathcal{L}_{\text{expert},k} = -\mathbb{E}_{(x,u^\star)\sim\mathcal{D}} \frac{1}{|\mathcal{I}_k^\star|} \sum_{t\in\mathcal{I}_k^\star} \log p_{\phi_k}(u_t^\star \mid x, u_{<t}^\star). \tag{2}$$

Each specialist supervises its stage content and delimiters, with Issue Filtering also covering the final issue list. Preceding stages remain in context; specialists are frozen during distillation.

4.2 Stage-Conditioned Distillation

For each input, the student generates a complete response $u \sim p_\theta(\cdot \mid x)$. Let $s(t)$ identify the stage containing token t. The final list uses the Issue Filtering teacher:

$$q_t(\cdot) = p_{\phi_{s(t)}}(\cdot \mid x, u_{<t}). \tag{3}$$

Teachers evaluate the *same* student-generated history: the Multidimensional Comparison specialist sees the student's bindings, and the Issue Filtering specialist sees its comparisons.

Explicit stage delimiters determine the assignment $s(t)$. We compare student and teacher probabilities on sampled tokens using a reverse-KL surrogate and normalize the loss within stage groups:

$$\mathcal{L}_{\text{stage}} = \mathbb{E}_{x,u}\left[\frac{\sum_{k\in\mathcal{G}(u)} \alpha_k \left(|\mathcal{I}_k(u)|^{-1} \sum_{t\in\mathcal{I}_k(u)} \widehat{d}_t\right)}{\sum_{k\in\mathcal{G}(u)} \alpha_k}\right]. \tag{4}$$

Here $\mathcal{I}_k(u)$ contains content tokens, $\mathcal{G}(u)$ indexes nonempty groups, and α_k gives their weights. Empty group sets contribute zero. Normalization balances stages of unequal length. Binding and comparison have weight 1; filtering and final output have weight 0.25. Replay supervises delimiters. Appendix B defines $\widehat{d}_t$ and clipping.

4.3 Structure and Decision Supervision

Annotated replay supplies separately normalized content, boundary, and decision supervision. The final issue-presence token belongs to the content term, while difference and event presence inside the

reasoning trace receive the separate decision term. Final report-presence and intermediate diagnostic supervision can therefore be weighted independently. We combine these terms as

$$\mathcal{L} = (1-\rho)\mathcal{L}_{\text{stage}} + \rho\big(\lambda_c\mathcal{L}_{\text{content}} + \lambda_b\mathcal{L}_{\text{boundary}} + \lambda_d\mathcal{L}_{\text{decision}}\big), \tag{5}$$

Here ρ is the replay fraction; we use unit coefficients $\lambda_c = \lambda_b = \lambda_d = 1$. Replay losses are averaged over annotated traces. Content supervision preserves annotated reasoning and output; boundary supervision preserves stage transitions; decision supervision emphasizes the presence of observed differences, reportable events, and prompt-permitted changes. Empty and nonempty decisions are balanced within each annotated trace.

Supervision allocation. Replay weights the reasoning stages by $0.5, 1, 1.5$ and assigns an additional weight E to nonempty comparison evidence, since unchanged attributes constitute 60.4% of annotated comparison blocks. The final list receives a separately normalized loss.

Decision supervision targets the first token distinguishing empty from nonempty difference or event lists, together with annotations distinguishing requested changes from unexpected differences. The selected configuration uses $E = 1.5$ and final issue-presence coefficient $g = 0.5$ in $\mathcal{L}_{\text{content}}$; Section 5.3 examines the completed component comparisons and sensitivity to these weights.

5 Experiments

5.1 Experimental Setup

Data and metrics. We train on 2,160 compact traces and evaluate the trained models on the reviewed benchmark. We report issue-detection precision, recall, and F_1 under the matching and pooling protocol in Section 3.4. Training retains subject assignments, paired evidence, empty comparisons, and prompt-permitted changes. All 9B distillation comparisons start from the same compact-trace SFT checkpoint. Evaluation scores the final issue list by semantic matching rather than requiring identical wording.

Models and baselines. The student and specialists use Qwen3.5-9B (Qwen Team, 2026a). Table 2 includes foundation models, supervised baselines, and on-policy distillation methods. The 9B SFT baseline learns complete compact traces and produces a diagnosis in one call. To test whether inference-time decomposition alone recovers the benefit of our pipeline, *SFT chain* reuses this exact checkpoint in three sequential calls: Subject Binding, Multidimensional Comparison, and Issue Filtering. Each call receives the preceding generated stages, with output limits of 2,048, 8,192, and 4,096 tokens. Its comparison blocks are generated within one causal call, whereas annotation uses parallel dimension-specific calls. It mimics the pipeline's diagnostic decomposition with shared weights, whereas our specialists receive separate stage-specific training. Qwen3.6-27B SFT provides a larger-model reference and the teacher for plain OPD (Agarwal et al., 2024).

The distillation baselines differ in how they construct and apply feedback. Video-OPD provides dense sampled-token teacher feedback; we use its base objective without task-specific video difficulty filtering (Li et al., 2026). Veto uses a forward-KL objective with a detached geometric teacher–student target (Jang et al., 2026). ReOPD replays a reference prefix and treats generation of the next diagnostic stage as an action (Liao et al., 2026). SDPO uses feedback-conditioned self-distillation, with feedback adapted to the reference traces (Hübotter et al., 2026); OPSD supplies privileged reference-trace context to the teacher (Zhao et al., 2026). These task adaptations retain the diagnostic representation and evaluation protocol, allowing us to compare alternative sources of feedback on the same final prediction task.

Implementation. SFT uses two epochs and 540 updates. The visual encoder is frozen while the language model and visual alignment layers are updated. Each specialist learns its assigned spans with preceding annotated stages in context; specialist parameters remain fixed during transfer. Stage OPD trains for up to 70 updates with 12 on-policy and four annotated trajectories per batch. The selected configuration uses learning rate 10^{-6}, evidence weight $E = 1.5$, and final issue-presence weight $g = 0.5$. Within each configuration, checkpoints are selected by peak monitored F_1. All metrics in a row come from the same checkpoint. Trained models share fixed media, prompts, and judging, with temperature 0.7, top-p 0.8, and top-k 20. Appendix D gives the remaining settings.

Table 2: Reference-conditioned video diagnosis (%). Post-trained results pool three decodings. Bold marks the best post-trained 9B score in each column.

Group	Model / method	Size	Precision	Recall	F_1
Closed-source	Gemini-3.5-Flash (Google DeepMind, 2026)	—	58.33	28.57	38.36
	GPT-5.5 (OpenAI, 2026a)	—	65.79	51.02	57.47
	GPT-5.6-Sol (OpenAI, 2026b)	—	41.18	57.14	47.86
	Qwen3.7-Plus (Alibaba Cloud, 2026a)	—	78.57	22.45	34.92
	Qwen3.7-Max (Qwen Team, 2026d)	—	66.67	16.33	26.23
	Qwen3.8-Max (Alibaba Cloud, 2026b)	—	58.33	42.86	49.41
Open-source	Qwen3.6-27B (Qwen Team, 2026b)	27B	65.38	34.69	45.33
	Qwen3.6-35B-A3B (Qwen Team, 2026c)	35B-A3B	61.11	22.45	32.84
	Qwen3.5-9B (Qwen Team, 2026a)	9B	48.15	17.69	25.87
	Qwen3.6-27B SFT	27B	49.54	36.73	42.19
	Qwen3.5-9B SFT (chain)	9B	47.27	**35.37**	40.47
	Qwen3.5-9B SFT	9B	51.22	28.57	36.68
	+ OPD (27B teacher) (Agarwal et al., 2024)	9B	60.81	30.61	40.72
	+ Video-OPD (Li et al., 2026)	9B	62.50	30.61	41.10
	+ Veto (Jang et al., 2026)	9B	**74.07**	27.21	39.80
	+ ReOPD (Liao et al., 2026)	9B	53.85	28.57	37.33
	+ SDPO (Hübotter et al., 2026)	9B	48.78	27.21	34.93
	+ OPSD (Zhao et al., 2026)	9B	39.53	23.13	29.18
	+ **Stage OPD**	9B	56.32	33.33	**41.88**

Evaluation controls. Three decoding seeds generate separate diagnoses from fixed weights. Their matching counts are pooled before computing precision, recall, and F_1. All comparisons use the same reviewed labels and semantic judge; duplicate reports cannot create additional matches.

5.2 Main Results

Diagnostic performance. Stage OPD reaches 41.88% F_1, compared with 36.68% for 9B SFT (Table 2). The 5.20-point improvement combines a precision increase from 51.22% to 56.32% with a recall increase from 28.57% to 33.33%. This is a 14.2% relative F_1 gain. The student recovers more reviewed defects while improving the reliability of its reports, showing that its coverage gain is not obtained simply by lowering the reporting threshold.

Stage OPD attains the highest F_1 among the compared 9B methods, exceeding Video-OPD's 41.10% and OPD with a 27B teacher at 40.72%. It comes within 0.31 points of the 27B SFT reference while using one-third as many model parameters at deployment, and improves precision by 6.78 points. The 27B model retains higher recall, while the student offers a competitive detection balance within a single 9B model. The precision–recall profiles distinguish the 9B alternatives. Veto has the highest precision but only 27.21% recall; Video-OPD reaches 30.61% recall. Stage OPD recovers more reviewed defects than these alternatives while retaining 56.32% precision, giving the strongest overall F_1 in this group. GPT-5.5 remains ahead at 57.47% F_1 in one zero-shot call, leading mainly through higher recall (51.02% versus 33.33%); Qwen3.8-Max and GPT-5.6-Sol reach 49.41% and 47.86%. This gap is expected: these models are far larger, and Qwen3.8-Max itself served as an annotation source, while the student already exceeds its other source, Gemini-3.5-Flash (38.36%). In the deployable regime we optimize for, training lifts the 9B base from 25.87% to 41.88%, closing about half of its gap to GPT-5.5 in a single self-contained generation.

Inference decomposition. Splitting 9B SFT into three calls raises F_1 from 36.68% to 40.47% with identical weights. The chain recovers additional defects, reaching 35.37% recall and 47.27% precision. Stage OPD improves F_1 by 1.41 points and precision by 9.05 points over the chain, while consolidating three calls into one. Across the three decodings, the chain produces 52 true positives and 58 false positives; Stage OPD produces 49 and 38, respectively. It retains most of the chain's defect coverage while eliminating 20 false reports, a 34.5% reduction.

The chain establishes a strong inference-time baseline without changing the SFT weights. Surpassing it in precision and F_1 with one call supports transferring the diagnostic process into the student's parameters. Stage OPD provides intermediate feedback on student-generated histories, teaching how the stages interact as well as their order. Stage OPD removes two intermediate model calls while improving diagnosis; its specialists are needed only during training.

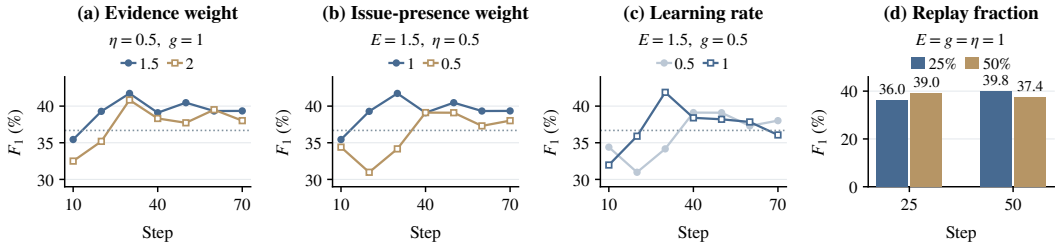

Figure 4: Ablations of (a) evidence weighting, (b) final issue-presence weighting, (c) learning rate, and (d) annotated replay. Settings shown above each panel are held fixed within that comparison; η is measured in 10^{-6}. Dotted horizontal lines mark 9B SFT.

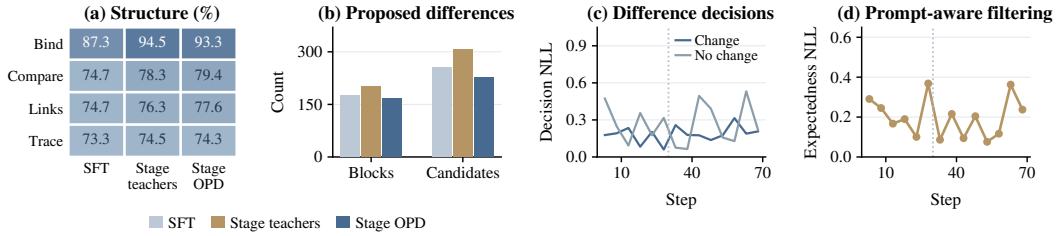

Figure 5: Teacher outputs and student decision learning. (a,b) Trace structure and candidate differences across 150 case–seed pairs. Links: candidate-to-event correspondence; Trace: complete validity. (c,d) Diagnostic-decision NLL; dotted lines mark the selected checkpoint.

5.3 Ablation Studies

Stage teachers and evidence weighting. Table 3 compares teacher organization and supervision weights. Plain OPD with a 27B teacher reaches 40.72% F_1, and its stabilized control reaches 40.14%. Stage teachers under the same stabilized recipe and uniform evidence weighting reach 40.91%. Increasing the evidence weight to $E = 2$ gives 41.45%; the selected $E = 1.5$, $g = 0.5$, $\eta = 10^{-6}$ setting reaches 41.88%. These comparisons support specialized feedback and emphasis on evidence-bearing attributes. Unchanged-attribute blocks remain useful supervision, but their frequency motivates balancing their contribution against explicit diagnostic evidence.

Table 3: Component ablations. Stab.: stage normalization and replay; $\eta = 10^{-6}$.

Teachers	Stab.	E	g	F_1
27B	—	—	—	40.72
27B	✓	1	1	40.14
27B	✓	2	1	39.45
27B	✓	1.5	0.5	39.98
Stage	—	—	—	40.86
Stage	✓	1	1	40.91
Stage	✓	2	1	41.45
Stage	✓	**1.5**	**0.5**	**41.88**

Hyperparameter sensitivity. Figure 4 examines evidence weight, final issue-presence weight, learning rate, and replay fraction. The comparisons separate evidence emphasis from final issue-presence supervision and examine how replay complements student-generated histories. We retain 25% replay to balance annotated supervision with student-generated trajectories. Evidence weighting changes the contribution of comparison content, while issue-presence weighting controls final reporting supervision. The learning-rate and replay panels show how optimization and the balance of annotated and generated trajectories affect detection under their respective fixed settings.

5.4 Teacher and Student Analysis

Teacher outputs. Figure 5a,b summarizes the specialists, SFT, and the selected student. The specialists provide stronger stage capabilities and broader candidate evidence for integration. Candidate observations may describe permitted changes or require consolidation. Proposal counts therefore characterize evidence available for integration, while precision and recall assess which observations become correct reports.

Diagnostic-decision learning. Figures 5c,d track decisions that connect visual comparison to reporting. From updates 1–10 to 21–30, unchanged-attribute NLL falls from 0.351 to 0.255 and expectedness NLL from 0.268 to 0.234. Change-presence NLL also falls, from 0.185 to 0.132 over

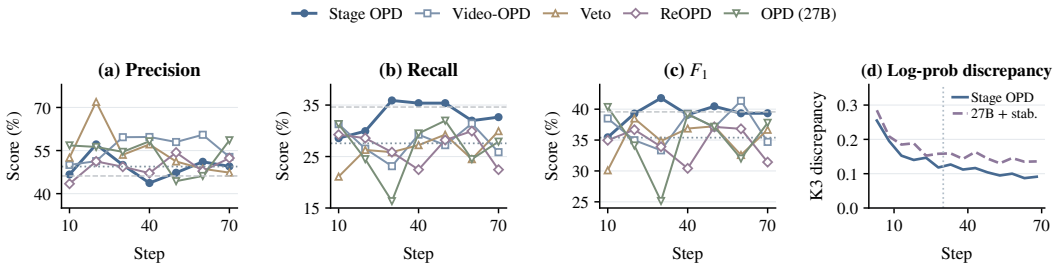

Figure 6: Training dynamics. (a–c) Precision, recall, and F_1. Dotted and dashed horizontal lines: SFT and SFT chain. (d) Stage-normalized K3 discrepancy in five-update windows; vertical line: selected update 30.

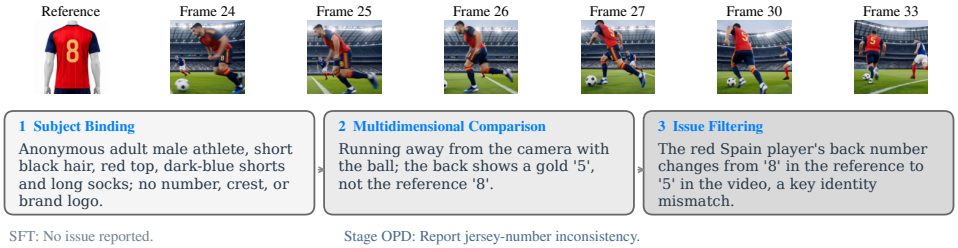

Figure 7: Shirt-number diagnosis across six sampled frames.

the same windows. These replay-target diagnostics show improved prediction of consistent attributes, differences, and prompt-permitted changes. They describe learning of the supervised decisions, while reviewed detection quality remains the criterion for selecting the checkpoint.

Performance throughout training. Figure 6 compares the monitored Stage OPD run with Video-OPD, Veto, ReOPD, and OPD using a 27B teacher. Jointly tracking precision and recall distinguishes broader defect coverage from more selective reporting. The checkpoint curves fluctuate across training, showing why an individual precision peak is insufficient for choosing the final model. Precision-focused alternatives such as Veto recover fewer reviewed defects, illustrating the need to assess reporting reliability together with defect coverage. SFT and SFT chain provide fixed reference points for the two inference procedures.

Teacher–student alignment. Figure 6d shows Stage OPD's stage-normalized K3 discrepancy falling from 0.255 over updates 1–5 to 0.091 over updates 66–70. The decreasing discrepancy characterizes optimization toward the stage teachers; reviewed detection performance provides the separate criterion for checkpoint selection.

Qualitative example. In Figure 7, Stage OPD identifies a shirt number changing from 8 to 5, whereas SFT reports no issue. The saved trace binds the player to ref_003, records the number mismatch as c002, and retains c002 as the reportable event. The stage excerpts connect the intermediate comparison to the final diagnosis.

6 Conclusion

Stage OPD integrates a decomposed diagnostic process into one model. Multipass annotation supplies evidence-linked traces that connect subject binding, multidimensional comparison, and issue filtering. Stage-specific teachers then supervise student-generated histories, while annotated replay reinforces descriptive content, stage transitions, and diagnostic decisions. This combination transfers specialized feedback into a unified student that performs the complete diagnosis in one generation. On the reviewed benchmark, the resulting 9B model improves both precision and recall over SFT, raising issue-detection $\boldsymbol{F_1}$ from 36.68% to 41.88% and approaching the 27B reference within 0.31 points. It also exceeds the three-call SFT chain, reducing false reports while consolidating inference into a single call. These results support using decomposition to organize supervision during training and integrating the learned capabilities at inference.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and the retrieval and discovery of related literature. We reviewed all AI-assisted revisions to ensure that they accurately reflected the intended meaning and verified the suggested references against their original sources.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, 2024.

Alibaba Cloud. Qwen3.7-Plus: Multimodal Agent Intelligence. Official Model Release, 2026a.

Alibaba Cloud. Alibaba Unveils Qwen3.8-Max: Its Largest and Most Capable Flagship Model to Date. Official Model Release, 2026b.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, 2015.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models. arXiv Preprint arXiv:2410.10818, 2024.

Zhuowei Chen, Bingchuan Li, Tianxiang Ma, Lijie Liu, Mingcong Liu, Yi Zhang, Gen Li, Xinghui Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom-Data: Towards a General Subject-Consistent Video Generation Dataset. arXiv Preprint arXiv:2506.18851, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Google DeepMind. Gemini 3.5: Frontier Intelligence with Action. Official Model Release, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, 2024.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li jia Li, and Yongxin Ni. Video-Bench: Human-Aligned Video Generation Benchmark. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. VideoScore2: Think before You Score in Generative Video Evaluation. arXiv Preprint arXiv:2509.22799, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. arXiv Preprint arXiv:1503.02531, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander 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*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation. arXiv Preprint arXiv:2601.20802, 2026.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable On-Policy Distillation through Adaptive Target Reformulation. In *Findings of the Association for Computational Linguistics*, 2026.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. In *International Conference on Learning Representations*, 2023.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-OPD: Efficient Post-Training of Multimodal Large Language Models for Temporal Video Grounding via On-Policy Distillation. arXiv Preprint arXiv:2602.02994, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-Turn On-Policy Distillation with Prefix Replay. arXiv Preprint arXiv:2607.04763, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri 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.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving Video Generation with Human Feedback. arXiv Preprint arXiv:2501.13918, 2025a.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-Consistent Video Generation via Cross-Modal Alignment. In *International Conference on Computer Vision*, 2025b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do Video LLMs Really Understand Videos? In *Findings of the Association for Computational Linguistics*, 2024.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-Teacher On-Policy Distillation for Capability Integration in LLM Post-Training. arXiv Preprint arXiv:2606.30406, 2026.

OpenAI. Introducing GPT-5.5. Official Model Release, 2026a.

OpenAI. GPT-5.6: Frontier Intelligence That Scales with Your Ambition. Official Model Release, 2026b.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. Official Model Release, 2026a.

Qwen Team. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model. Official Model Release, 2026b.

Qwen Team. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All. Official Model Release, 2026c.

Qwen Team. Qwen3.7: The Agent Frontier. Official Model Release, 2026d.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 627–635, 2011.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems*, 2024.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. OpenS2V-Nexus: A Detailed Benchmark and Million-Scale Dataset for Subject-to-Video Generation. In *Advances in Neural Information Processing Systems*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research*, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. arXiv Preprint arXiv:2601.18734, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. arXiv Preprint arXiv:2503.21755, 2025.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *International Conference on Learning Representations*, 2023.

Table 4: Information retained in compact traces.

Stage	Information retained in a compact trace
Subject Binding	Subject type and identity, reference assignment, distinguishing visual attributes.
Multidimensional Comparison	Subject and attribute dimension, reference evidence, video evidence, observed differences and expectedness annotations.
Issue Filtering	Event membership, subject assignment, source-change identifiers, expectedness, and reasons to report or suppress differences.
Final output	The consolidated issue list, including an explicit empty list when no issue is reportable.

A Training Data and Annotation

Annotation and trace construction. Gemini-3.5-Flash and Qwen3.8-Max annotate subject bindings and parallel dimension-wise comparisons, including expectedness. Filtering recorded comparisons and text-only event consolidation produce reportable events. The compact-source collection combines 1,120 Gemini-derived and 1,068 Qwen-derived examples. Filtering removes 22 examples above 65,536 multimodal prompt tokens; removing the last six aligns the source with the post-training batch size, yielding 2,160 traces. Annotation uses nominal four-frame-per-second sampling subject to frame caps; all subsequent stages reuse the saved frames.

A text-only rewriter shortens identities and paired evidence while preserving the source judgments. It groups descriptions of the same continuous event: each reportable source change belongs to one group, with its subject and evidence links preserved. The grouped event retains the source annotations and a source-supported explanation. Known evaluation identifiers and human-edited training cases are excluded. Table 4 summarizes the retained information.

Coverage. The retained collection contains 8,485 change records and 15,117 comparison blocks, of which 9,137 contain no differences. Thus, even issue-positive videos supply many unchanged-attribute examples. Figure 8 shows comparison-block composition and the distribution of subject categories. A trace may contain several subject categories, so category totals exceed the number of traces. The annotations are machine-generated training targets; the reviewed benchmark described below provides the evaluation labels.

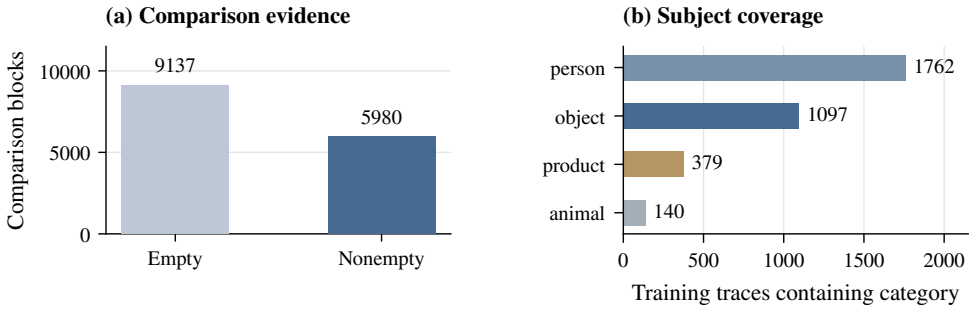

Figure 8: Training-data composition. Subject categories can overlap.

B Objective and Optimization

Sampled-token discrepancy. For the student-generated token u_t, define the student–teacher log-probability ratio

$$\ell_t = \log p_\theta(u_t \mid x, u_{<t}) - \log q_t(u_t). \tag{6}$$

We use the nonnegative reverse-KL estimator

$$\widehat{d}_t = \exp(-\ell_t) + \ell_t - 1. \tag{7}$$

For a token sampled from $p_\theta(\cdot \mid x, u_{<t})$ at a fixed prefix, the unclipped expectation equals $D_{\mathrm{KL}}(p_\theta \| q_t)$. The implemented loss substitutes $\widetilde{\ell}_t = \mathrm{clip}(\ell_t, -20, 20)$ into Eq. 7 and caps the

result at 10; this clipped surrogate is used in Eq. 4. Teacher scores are detached. Gradients pass through sampled student log probabilities; the surrogate does not include the full sequence-level sampling-distribution gradient.

Subject Binding, Multidimensional Comparison, Issue Filtering, and final-output groups have weights $1, 1, 0.25, 0.25$, respectively. Each is token-normalized, then combined using normalized weights over present groups. Stage delimiters are supervised only through replay.

Replay supervision. Replay reasoning content uses stage weights $0.5, 1, 1.5$ for binding, comparison, and reporting. If $\mathcal{I}_R$ contains reasoning tokens and w_t is the stage weight, the reasoning loss is

$$\mathcal{L}_R = -\frac{\sum_{t\in\mathcal{I}_R} w_t e_t \log p_\theta(u_t^\star \mid x, u_{<t}^\star)}{\sum_{t\in\mathcal{I}_R} w_t e_t}, \qquad e_t = \begin{cases} E & \text{eligible comparison evidence,} \\ 1 & \text{otherwise.} \end{cases} \tag{8}$$

The content loss is

$$\mathcal{L}_{\text{content}} = \mathbb{E}_{(x,u^\star)\sim\mathcal{D}}[\mathcal{L}_R + \mathcal{L}_{\text{final}} + g\mathcal{L}_{\text{presence}}]. \tag{9}$$

Here $\mathcal{L}_{\text{final}}$ averages cross-entropy over final-output tokens excluding the separately normalized issue-presence token, whose loss is $\mathcal{L}_{\text{presence}}$. Thus g controls this content-loss component; λ_d controls the distinct diagnostic-decision loss. Eligible evidence consists of nonempty comparison content excluding decision-marked tokens; those tokens keep their base stage weight in $\mathcal{L}_R$. The selected model uses $E = 1.5$ and $g = 0.5$. Boundary loss averages cross-entropy over stage delimiters. Decision supervision includes difference presence, reportable-event presence, and prompt-aware expectedness judgments. Empty and nonempty outcomes receive equal weight when both are present within a trace; difference presence receives twice the weight of each other decision family. Presence supervision is placed at the first token that distinguishes empty from nonempty serialization. The default content, boundary, and decision coefficients in Eq. 5 are all one, and the replay fraction is $\rho = 0.25$.

Student and specialist initialization. 9B SFT uses stage-weighted reasoning and separate final-output and issue-presence losses. Specialists supervise their assigned spans and delimiters; the Issue Filtering specialist also supervises final serialization. Preceding annotated stages remain in context. Training freezes the visual encoder and updates the language model and visual alignment layers.

C Evaluation Protocol

Matching and metrics. A frozen text-only Qwen3.7-Max judge matches predicted and reviewed issues by subject and defect semantics. Fixed one-to-one assignment prevents duplicate reports from receiving extra credit. Precision, recall, and F_1 pool matching counts over three decodings, using the same reviewed labels and judging prompts for every model. Each prediction and reviewed issue can participate in at most one match. Three generations give 150 case–seed and 147 issue–seed observations. Native thinking is disabled for the externally structured trace.

Checkpoint selection. Each configuration uses the checkpoint with the highest monitored F_1. All metrics in a row come from the same checkpoint; the three seeds vary decoding.

Trace diagnostics. The same checker evaluates subject bindings, comparison schemas, candidate-to-event links, and complete traces on 150 matched case–seed pairs. Teacher traces are generated sequentially by the three specialists. Candidate counts describe proposals before filtering, while the final detection metrics measure semantic correctness.

D Training Settings

Table 5 gives the shared settings. The selected Stage OPD checkpoint is update 30, with $E = 1.5$, $g = 0.5$, and learning rate 10^{-6}.

Table 5: Shared training and evaluation settings.

Setting	Value
Student / stage specialists	Qwen3.5-9B / three Qwen3.5-9B models
Larger SFT baseline	Qwen3.6-27B
SFT data and schedule	2,160 traces; two epochs; 540 updates; batch size 8
SFT learning rate	Cosine decay from 10^{-6} to 10^{-7}; warmup fraction 0.03
SFT training seed	42
Transfer schedule	70 updates; 12 on-policy traces and 4 replay traces per batch
Transfer learning rates	2.5×10^{-7}, 5×10^{-7}, 10^{-6}
Evidence weights in the study	$E = 1$, $E = 1.5$ (main), and $E = 2$
Evaluation context / output limits	73,728 / 32,768 tokens
Image / video-frame token caps	1,024 / 768
On-policy sampling	Temperature 1; no top-k or nucleus truncation
Evaluation decoding	Temperature 0.7; top-p 0.8; top-k 20
Presence / repetition penalties	1.5 / 1.05
Decoding seeds	1234, 2234, 3234

Component comparisons. Table 3 compares teacher organization and supervision weights. The selected joint setting uses $E = 1.5$, $g = 0.5$, and $\eta = 10^{-6}$. Figure 4 varies evidence weighting, issue-presence weighting, learning rate, and replay fraction under the settings shown above each panel.

Decoding variability. For decoding seeds 1234, 2234, and 3234, Stage OPD obtains 40.96, 40.86, and 43.48% F_1 on the benchmark; 9B SFT obtains 43.90, 32.50, and 29.63%. These are repeated generations with fixed model weights, not independent training runs. Main-table scores use pooled prediction and matching counts.

E Training Dynamics

Decision NLL in Figure 5 uses the selected $E = 1.5$, $g = 0.5$, $\eta = 10^{-6}$ run, pooling applicable replay tokens in five-update windows. The main text pools two adjacent windows into ten-update summaries. Figure 6a–c reports checkpoint performance at updates 10–70, with SFT and SFT chain as horizontal references.

Figure 6d shows the logged clipped K3 discrepancy over all 70 updates in five-update windows. Stage OPD and the stabilized 27B control use stage-weighted means divided by the on-policy fraction. Each curve uses its own training rollouts. Video-OPD, ReOPD, Veto, and plain OPD are compared through detection performance in panels (a–c), without substituting their differently normalized training objectives into panel (d).

F Prompts and Example

The following English summaries describe the canonical prompt templates and output contracts. They are abridged for readability; the complete Chinese prompt template (including variants for specific dimensions) will be made open-source upon the paper's acceptance. Binding and comparison use vision–language calls. Parallel dispatch and filtering operate on the staged outputs; the filtering summary describes how comparisons feed the retained annotations and final report.

F.1 Shared Inputs and Prompt Templates

Each call receives the generation prompt, numbered reference images, and the saved chronological frame sequence. A reference's position is bound to a logical identifier such as `ref_000`; these identifiers retain the same meaning throughout the pipeline. The comparison stage reuses the binding result instead of assigning identities independently for every attribute.

Subject Binding: identify the comparison targets

Identify the visible subjects and associate them with the permitted reference identifiers. For each subject, record its type, name, primary status, reference assignment, and distinguishing appearance. Use identity cues that remain meaningful across poses and views. A subject's temporary absence or occlusion is not an appearance defect. Return structured bindings and any unassigned references; do not invent a reference for an unmatched subject.

The binding result determines the comparison calls. Each primary subject is examined along shape and physical structure, color and material, and appearance for people or animals (brand markings and text for products). Secondary subjects are handled together. The dimension calls run independently after binding and are joined before filtering. Assigned reference images provide positive evidence; another subject's references serve only as negative identity anchors.

Multidimensional Comparison: collect paired evidence

Compare `{SUBJECT}` along `{DIMENSION}` against its assigned references. Describe the relevant reference attributes and the visible video evidence before listing differences. Compare only observable regions; viewpoint, pose, occlusion, camera motion, and lighting alone do not establish a defect. Use `differences: []` when the attribute is consistent.
For each difference, record its identifier, attribute, paired evidence, expectedness, and reason. Severity scores in $[0, 1]$ map to mild (轻度) below 0.5 and severe (重度) otherwise. These labels supplement the semantic description of the defect. Prompt-permitted transformations and clothing changes are excluded from consistency issues.

Filtering and compact rewriting: preserve the evidence links

Remove prompt-permitted differences and collect excluded observations in `ignored_changes`. Consolidate retained comparisons concerning the same subject and defect. Preserve source-change identifiers and source-supported explanations so each report remains linked to its evidence. Do not invent differences during rewriting. Emit an empty event list when no issue is reportable.

The resulting training target serializes bindings, dimension comparisons, and filtering in order, followed by the final `issues` object. Rewriting shortens descriptions but does not replace an expectedness judgment, reverse a reference assignment, or invent a defect. This creates a compact trace in which every final issue can be connected back to comparison evidence.

F.2 Worked Annotation Example

Figure 2 uses training row 6, which contains three reference images and 60 sampled frames. The generation prompt describes Ava in bed wearing black nightwear and speaking by phone with Chloe, who is seated on her sofa. The saved bindings associate Ava with `ref_002` and Chloe with `ref_000`. The examples below retain original field values and identifiers; ellipses omit other records.

Structured trace: selected original fields

```
<stage name="subject_identification">
{"subject_name":"Ava","reference_ids":["ref_002"]}
{"subject_name":"Chloe","reference_ids":["ref_000"]}
</stage>
<stage name="dimension_comparison" subject="Chloe"
      dimension="外观特征">
"differences":  [{"change_id":"c005","attribute":"发型",
  "reference_value":"黑色头发扎成高盘发发髻",
  "video_value":"黑色长发披散在肩头，未盘起",
  "is_expected":false, ...}]
</stage>
<stage name="dimension_comparison" subject="Ava"
      dimension="颜色与材质">
"differences":  [{"change_id":"c002","attribute":"服装",
  "is_expected":true, ...}]
```

```
</stage>
<stage name="dimension_comparison" subject="Ava"
      dimension="外观特征">
"differences":  []
</stage>
[...  other comparisons and fields omitted ...]
<stage name="change_report">
"ignored_changes":  [{"change_id":"c002",
  "disposition":"expected"}, ...]
"events":  [{"event_id":"e001","subject_name":"Chloe",
  "source_change_ids":["c003","c004","c005"], ...}]
[...  event links and descriptions omitted ...]
</stage>
```

From bindings to comparisons. The reference assignment keeps the clothing and hairstyle observations attached to the correct person. Ava's shape/structure comparison records the transition from a jacket to a camisole as `c001`; the color/material comparison records the corresponding change from a fur collar to lace as `c002`. These parallel observations concern related aspects of the same requested outfit. An empty appearance comparison supplies complementary evidence that the corresponding attributes remain consistent.

From comparisons to filtering. Ava's clothing changes are marked expected and appear in `ignored_changes`. Chloe's appearance comparison instead records `c005`: her reference hairstyle is a high bun, whereas the video shows her hair loose around her shoulders. The source marks this change unexpected. Filtering retains its subject, visual evidence, and source identifier in the reportable event. The final report therefore includes Chloe's hairstyle issue while suppressing Ava's requested clothing differences.

What the example illustrates. The compact trace preserves both positive and negative supervision: shared bindings, paired evidence, empty comparisons, excluded changes, and a reportable event. A final issue list alone would omit why Ava's visible difference is permitted and how Chloe's event is grounded. The intermediate fields make these decisions available to the student and its stage specialists while preserving the original annotation. The supplementary material provides the complete prompt templates used to construct these annotations.