FLUX-OPD: ON-POLICY DISTILLATION WITH EVOLVING CONTEXTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model training in open-ended domains lacks verifiable rewards, making task preferences difficult to formalize as effective supervision. Contexts derived from task feedback can provide richer supervision, yet fixed contexts cannot adapt as the student evolves, motivating contexts that evolve with student performance. However, directly using evolving contexts as in-training supervision results in unstable distillation targets and conflicting distributions, requiring mechanisms to stabilize the target and downweight conflicts. In this paper, we characterize context distillation through a reverse-KL decomposition, showing that the student is distilled toward the geometric mean of context-conditioned teachers, while the objective contains a conflict term measuring disagreement among these teachers. Motivated by this characterization, we propose Flux-OPD, an OPD paradigm that uses evolving contexts as in-training supervision to incorporate contextual signals derived from task feedback. Flux-OPD treats differences between context-conditioned and context-free teachers as contextual difference signals, injects them as contextual corrections into the context-free teacher, and weights the correction strength using the conflict term. Experiments on open-ended tasks show that Flux-OPD achieves the best reported aggregate results among the evaluated OPD paradigms, highlighting the potential of combining teacher supervision with evolving contexts as in-training supervision.

1 INTRODUCTION

Large language models achieve remarkable progress through Reinforcement Learning with Verifiable Rewards in verifiable domains (Shao et al., 2024). However, open-ended domains often lack verifiable rewards, making it difficult to formalize supervision signals that effectively capture task preferences. Rubrics-as-Rewards (RaR) (Gunjal et al., 2025; Xiao et al., 2026) uses rubrics as supervision but still provides sparse feedback via a scalar score, limiting its effectiveness in capturing complex preference priors (Ye et al., 2026a; Lu & Lab, 2025; Schulman & Lab, 2025).

An alternative is to use contexts as supervision signals. Context Distillation and On-Policy Context Distillation (OPCD) (Askell et al., 2021; Ye et al., 2026c) distill the student toward a context-conditioned teacher, enabling it not only to benefit from dense teacher supervision, as in vanilla On-Policy Distillation (OPD) (Agarwal et al., 2024), but also to directly learn behaviors that the teacher exhibits after observing the contexts. Thus, explicit contexts can provide additional task-related supervision that is difficult to formalize in open-ended domains. However, we find that directly distilling toward the context-conditioned teacher (OPCD) underperforms distilling toward the context-free teacher (OPD), as shown in Fig. 1. Since these contexts are extracted before training, they remain fixed as the student evolves. This motivates us to **explore contexts that evolve with student performance**.

The key challenge is to incorporate evolving contexts as supervision signals throughout training without disrupting optimization. Online Experiential Learning (OEL) (Ye et al., 2026b) alternates between deployment and OPCD training, accumulating experience items from extensive student trajectories as contexts. To provide in-training supervision, we extend this paradigm to a single training run, using only a small set of trajectories at each update to provide timely feedback on the current student. However, as shown in Fig. 1, evolving contexts do not yield consistent gains. Each update can abruptly shift the target distribution, accompanied by loss surges and less stable training. More-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

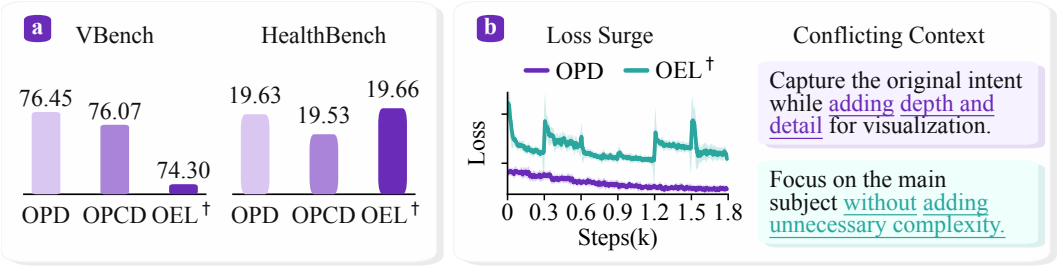

Figure 1: Motivations. (a) OPCD underperforms OPD, while OEL† does not yield consistent gains. (b) Context updates are accompanied by loss surges, while conflicts among contexts can produce conflicting distributions. † denotes our single-run adaptation of OEL.

over, conflicting contexts can produce conflicting target distributions. These observations motivate us to **stabilize the distillation target and downweight conflicting distributions**.

To better leverage contexts, we analyze their effect on the reverse KL objective in the OPCD paradigm. We show that this objective can be decomposed into *a distillation term* and *a conflict term*. The former computes the KL divergence between the student and the geometric mean of context-conditioned teacher distributions, while the latter measures conflicts among these teachers. Conditioned on fixed sampled histories, the decomposition shows that the conflict term contributes no direct gradient, leaving the distillation term as the directly optimized component of the objective.

Motivated by this characterization, we propose **Flux-OPD**, an OPD paradigm that uses evolving contexts as in-training supervision to incorporate contextual signals derived from task feedback. It is an iterative training process that divides a single training run into multiple iterations, each comprising context extraction and context distillation. The teacher extracts contexts from student trajectories during context extraction and conditions on these contexts during context distillation.

We design two distillation strategies: **contextual correction** and **contextual weighting**. To stabilize the distillation target, the contextual correction strategy anchors it to the relatively stable context-free teacher and injects contextual difference signals that capture task preferences from evolving contexts, rather than directly treating context-conditioned teachers as distillation targets. Contextual difference signals are defined as the differences between the context-conditioned and context-free teachers. To downweight conflicting distributions, the contextual weighting strategy adjusts the correction strength based on the conflict term, strengthening corrections when context-conditioned distributions are consistent and weakening them when they conflict.

We evaluate our method on two open-ended tasks with real-world applicability: prompt optimization for video generation and medical question answering, both of which involve complex preferences and rely on rich contextual information. Results show that Flux-OPD outperforms existing OPD paradigms on both tasks across different student-teacher settings. Our contributions are threefold:

- We explore evolving contexts as in-training supervision for on-policy distillation in open-ended domains. We identify two key challenges in directly incorporating such supervision: unstable distillation targets and conflicting context-conditioned distributions.
- We characterize context distillation through a reverse-KL decomposition into a geometric-mean target and a conflict term. This motivates Flux-OPD with contextual correction and weighting, combining stable teacher and evolving contextual supervision.
- Experiments on open-ended tasks show that Flux-OPD achieves the best reported performance among the evaluated OPD paradigms, highlighting the potential of combining teacher supervision with evolving contexts.

2 Related Work

2.1 Open-Ended Reward Modeling

Large language model training in open-ended domains lacks verifiable rewards, making task preferences difficult to formalize as effective supervision. Reinforcement Learning with Verifiable Re-

wards (RLVR) compresses these fine-grained preferences into scalar rewards. Rubrics-as-Rewards (RaR) (Gunjal et al., 2025; Xiao et al., 2026) uses rubrics as structured supervision but still provides sparse feedback. On-Policy Distillation (OPD) trains on trajectories sampled from the current student policy and provides dense teacher supervision (Agarwal et al., 2024; Zhao et al., 2026; Hübotter et al., 2026; Fu et al., 2026; Gu et al., 2024), but does not directly incorporate contextual signals derived from task feedback (Gudibande et al., 2023; Yang et al., 2026a;b). Flux-OPD complements teacher supervision with evolving contexts derived from task feedback as in-training supervision.

2.2 Context Distillation

Context Distillation (Askell et al., 2021; Snell et al., 2022) distills a student from a context-conditioned teacher, enabling it to learn behaviors that the teacher exhibits after observing the contexts. On-Policy Context Distillation (OPCD) (Ye et al., 2026c) further samples from the current student, reducing the mismatch between offline data and the student's generation distribution. Although contexts can provide additional task-related supervision in open-ended domains, contexts fixed before training cannot adapt to the evolving student. OEL (Ye et al., 2026b) alternates between deployment and OPCD training, updating contexts between runs. Directly extending this paradigm to a single continuous run results in an unstable distillation target and conflicting distributions. Flux-OPD instead uses evolving contexts as in-training supervision by distilling contextual difference signals into a context-free anchor rather than fully imitating context-conditioned teachers.

3 Preliminaries

3.1 Paradigms

In vanilla On-Policy Distillation (OPD), given a user prompt x, the student π_θ autoregressively generates a response $y = (y_1, \ldots, y_T)$. Let $h_t = (x, y_{<t})$ denote the decoding history and $\mathcal{V}$ the vocabulary. At step t, the student and teacher distributions are defined as

$$\begin{aligned} p_\theta(v \mid h_t) &= \pi_\theta(y_t = v \mid h_t), \\ q_0(v \mid h_t) &= \pi_{\mathrm{T}}(y_t = v \mid h_t), \quad v \in \mathcal{V}. \end{aligned} \tag{1}$$

In On-Policy Context Distillation (OPCD), the teacher additionally receives a privileged context $c \sim \mathcal{C}$, while the student observes only h_t. The context-conditioned teacher distribution is

$$q_c(v \mid h_t) = \pi_{\mathrm{T}}(y_t = v \mid h_t, c). \tag{2}$$

Training minimizes the reverse KL under on-policy histories:

$$\mathcal{L}(\theta) = \mathbb{E}_{h_t} \mathbb{E}_{c \sim \mathcal{C}} \left[D_{\mathrm{KL}} \left(p_\theta(\cdot \mid h_t) \| q_c(\cdot \mid h_t) \right) \right]. \tag{3}$$

For simplicity, fixing h_t and omitting the condition gives

$$\mathcal{L}(\theta) = \mathbb{E}_{c \sim \mathcal{C}} \left[D_{\mathrm{KL}}(p_\theta \| q_c) \right]. \tag{4}$$

3.2 KL Decomposition

For a fixed decoding history h_t at step t, let $p_\theta(v) = p_\theta(v \mid h_t)$ and $q_c(v) = q_c(v \mid h_t)$. We omit h_t for readability.

Forward KL Prior work (Yang et al., 2026a) decomposes the forward KL objective into a distillation term toward a marginal teacher and a mutual information term:

$$\mathbb{E}_c \left[D_{\mathrm{KL}}(q_c \| p_\theta) \right] = \underbrace{D_{\mathrm{KL}}(q_{\mathrm{mar}} \| p_\theta)}_{\text{distillation term}} + \mathbb{E}_c \left[D_{\mathrm{KL}}(q_c \| q_{\mathrm{mar}}) \right], \tag{5}$$

where the marginal teacher is defined as

$$q_{\mathrm{mar}}(v) = \mathbb{E}_c[q_c(v)]. \tag{6}$$

Thus, under forward KL, the student is distilled toward the **arithmetic mean** of context-conditioned teachers.

Reverse KL We characterize context distillation through the objective in Eq. (4).

Proposition 1 (Reverse KL Decomposition)**.** *For a fixed decoding history, let p_θ and $\{q_c\}$ be distributions over the same vocabulary $\mathcal{V}$, where $q_c(v) > 0$ for every context c and token $v \in \mathcal{V}$. Let $\mathbb{E}_c$ denote expectation with respect to a fixed context distribution. Define the normalized geometric mean of context-conditioned teachers and its normalization constant as*

$$q_{\text{geo}}(v) = \frac{\exp(\mathbb{E}_c[\log q_c(v)])}{Z}, Z = \sum_{v \in \mathcal{V}} \exp(\mathbb{E}_c[\log q_c(v)]) \,. \tag{7}$$

Then, the reverse KL objective decomposes as

$$\mathbb{E}_c \left[D_{\text{KL}}(p_\theta \| q_c) \right] = \underbrace{D_{\text{KL}}(p_\theta \| q_{\text{geo}})}_{\textit{distillation term}} + \underbrace{(-\log Z)}_{\textit{conflict term}}, \tag{8}$$

where $-\log Z \geq 0$.

Proof. Expanding the reverse KL objective gives

$$\mathcal{L}(\theta) = \mathbb{E}_c \left[\sum_{v \in \mathcal{V}} p_\theta(v) \log \frac{p_\theta(v)}{q_c(v)} \right] = \sum_{v \in \mathcal{V}} p_\theta(v) \log p_\theta(v) - \sum_{v \in \mathcal{V}} p_\theta(v) \mathbb{E}_c[\log q_c(v)]. \tag{9}$$

According to Eq.(7),

$$\mathbb{E}_c[\log q_c(v)] = \log q_{\text{geo}}(v) + \log Z. \tag{10}$$

Substituting this relation into Eq. (9) yields

$$\mathcal{L}(\theta) = \sum_{v \in \mathcal{V}} p_\theta(v) \log \frac{p_\theta(v)}{q_{\text{geo}}(v)} - \log Z \sum_{v \in \mathcal{V}} p_\theta(v) = D_{\text{KL}}(p_\theta \| q_{\text{geo}}) - \log Z, \tag{11}$$

which gives Eq. (8).

Moreover, by Jensen's inequality (Jensen, 1906),

$$Z = \sum_{v \in \mathcal{V}} \exp\left(\mathbb{E}_c[\log q_c(v)]\right) \leq \sum_{v \in \mathcal{V}} \mathbb{E}_c[q_c(v)] = 1, \tag{12}$$

Therefore, $-\log Z \geq 0$. □

Under reverse KL, the student is thus distilled toward the **normalized geometric mean** of context-conditioned teachers, rather than the arithmetic mean under forward KL. The geometric mean assigns high probability only to tokens that receive consistently high probability across contexts. Moreover, the conflict term $-\log Z$ reflects distributional disagreement among these teachers: lower distributional overlap decreases Z and increases $-\log Z$.

Corollary 1 (Gradient Independence of the Conflict Term)**.** *For fixed decoding histories, a fixed context distribution, and fixed context-conditioned teacher distributions, the conflict term $-\log Z$ is independent of the current student token distribution. Therefore, it contributes no direct gradient with respect to the student parameters through p_θ.*

As in standard OPD, token-level optimization treats sampled histories as fixed; under this convention, only the distillation term is directly optimized, with the geometric mean serving as the target. Applied across iterations, this fixed-history decomposition highlights target shifts and varying conflicts, motivating us to anchor the target to the context-free teacher and downweight corrections under high conflict. Fig. 2(a) visualizes the mean output embeddings of different distillation targets, showing that context-conditioned teachers undergo large shifts which may destabilize training, whereas the context-free teacher remains stable across updates. Moreover, as shown in Fig. 2(b), variations in the conflict value reflect changes in the sampled histories and context pools across iterations rather than direct optimization of the conflict term.

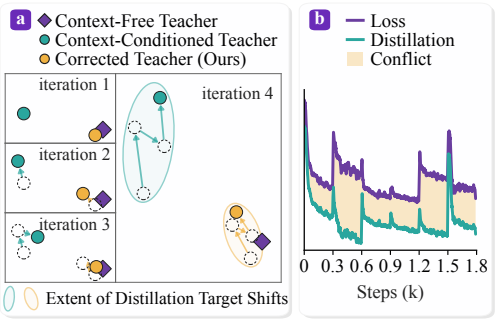

Figure 2: Characterizing training with evolving contexts using the reverse-KL decomposition. (a) Evolving contexts cause shifts in context-conditioned teacher distributions. (b) The conflict term varies with the sampled histories and contexts rather than through direct optimization.

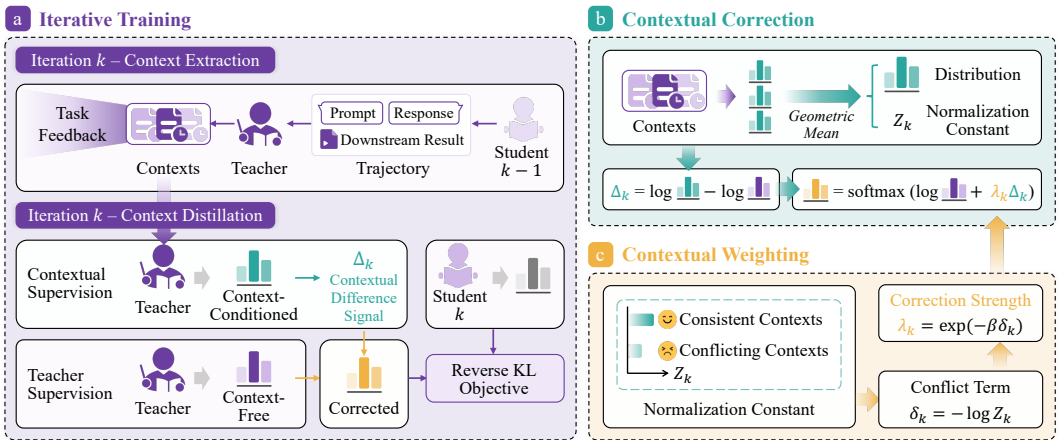

Figure 3: Overview of Flux-OPD. (a) illustrates the *iterative training* process, which divides a single training run into K iterations, each consisting of context extraction and context distillation. (b) illustrates the *contextual correction* strategy, which corrects the distillation target by combining the context-free teacher distribution with contextual difference signals from context-conditioned teachers. (c) illustrates the *contextual weighting* strategy, which further adjusts correction strengths using the conflict term, strengthening corrections when context-conditioned teacher distributions are consistent and weakening them when they conflict.

4 Method

We propose Flux-OPD, an OPD paradigm that uses evolving contexts as in-training supervision to incorporate contextual signals derived from task feedback. As shown in Fig. 3, it is an iterative training process including two distillation strategies: contextual correction and contextual weighting.

4.1 Iterative Training

To adapt contextual supervision to the evolving student, we update contexts with student performance. We update contexts within a single training run, using a small set of trajectories at each update to provide timely feedback on the current student. These evolving contexts make the training an iterative process that divides a single training run into K iterations, each consisting of context extraction and context distillation. As shown in Fig. 3(a), the teacher extracts contexts from student trajectories during context extraction and conditions on these contexts during context distillation.

Context Extraction Following existing context distillation methods (Ye et al., 2026c;b), contexts consist of experience items extracted from student interaction trajectories. At the beginning of iteration $k \in \{1, 2, \ldots, K\}$, we collect trajectories from the previous student policy $\pi_{\theta_{k-1}}$:

$$\mathcal{D}_{k-1} = \{(x_i, \hat{y}_i)\}_{i=1}^{n}, \tag{13}$$

where x_i is the user prompt, $\hat{y}_i$ is the student response and optional supplements, such as downstream video generated from $\hat{y}_i$ in prompt optimization task and rubrics in medical question answering task. For the first iteration, user prompts are randomly sampled from the entire training set. For subsequent iterations, they are randomly sampled from the training data seen in the previous iteration.

Given each trajectory, the teacher extracts new experience items that differ from existing ones and accumulate them with previous experience items to form a complete context. This process is repeated for M turns with different random seeds, yielding a context pool:

$$\mathcal{C}_k = \{c_{k,1}, \ldots, c_{k,M}\}. \tag{14}$$

Context Distillation Contexts sampled from the context pool $\mathcal{C}_k$ condition the teacher, allowing the target distribution to evolve with student performance. Since the context-free teacher provides a stable anchor for the base target, while the conflict term indicates conflicting context-conditioned teacher supervision, we stabilize distillation through contextual correction and contextual weighting.

4.2 CONTEXTUAL CORRECTION

To stabilize the distillation target, the contextual correction strategy anchors it to the context-free teacher and injects contextual difference signals derived from evolving contexts, rather than directly using context-conditioned teachers as targets, as shown in Fig. 3(b).

For readability, we suppress the dependence of $q_{\mathrm{geo},k}$, Z_k, δ_k, and λ_k on the decoding history h_t and position t. For iteration k, at each response position, we first compute the normalized geometric mean of context-conditioned teachers:

$$\begin{aligned} \log \tilde{q}_{\mathrm{geo},k}(v) &= \frac{1}{|\mathcal{C}_k|} \sum_{c \in \mathcal{C}_k} \log q_c(v), \\ q_{\mathrm{geo},k}(v) &= \frac{\exp(\log \tilde{q}_{\mathrm{geo},k}(v))}{Z_k}, \end{aligned} \tag{15}$$

where

$$Z_k = \sum_{v \in \mathcal{V}} \exp(\log \tilde{q}_{\mathrm{geo},k}(v)). \tag{16}$$

This is the empirical counterpart of Eq.(7), using a uniform average over the finite context pool $\mathcal{C}_k$. Since the context-conditioned teacher is constructed in log-probability space, we compute the contextual difference signal relative to the context-free teacher q_0, as

$$\Delta_k(v) = \log q_{\mathrm{geo},k}(v) - \log q_0(v). \tag{17}$$

The final corrected teacher distribution is then obtained through log-space interpolation:

$$q_k^{\mathrm{flux}}(v) = \mathrm{softmax}_v \left(\log q_0(v) + \lambda_k \Delta_k(v) \right), \tag{18}$$

where $\lambda_k \in [0, 1]$ controls the strength of contextual correction. When $\lambda_k = 0$, the target reduces to the context-free teacher q_0; when $\lambda_k = 1$, it recovers the normalized geometric-mean teacher $q_{\mathrm{geo},k}$. Thus, the context-free teacher q_0 provides a stable anchor, while the contextual difference signal Δ_k incorporates contextual information derived from task feedback.

4.3 CONTEXTUAL WEIGHTING

To downweight conflicting distributions, the contextual weighting strategy adjusts the correction strength using the conflict term, as shown in Fig. 3(c).

From the reverse KL decomposition in Eq.(8), the conflict term measures distributional disagreement among context-conditioned teachers. Motivated by this observation, we use it as a disagreement indicator to adjust the correction strength:

$$\delta_k = -\log Z_k, \tag{19}$$

where a larger δ_k indicates lower overlap among context-conditioned teacher distributions at the current decoding position.

We design the weighting mechanism from a conflict-aware perspective: corrections are strengthened when context-conditioned teacher distributions are consistent and weakened when they conflict. Specifically, we use a monotonically decreasing mapping:

$$\lambda_k = \exp(-\beta \delta_k), \tag{20}$$

where β is a scaling factor. At each decoding position, the same weight is applied to the contextual correction across all vocabulary logits.

4.4 TRAINING OBJECTIVE

In summary, Flux-OPD trains the student toward the reformulated target at each iteration by computing the reverse KL divergence between the student distribution and the corrected teacher distribution:

$$\mathcal{L}_k^{\mathrm{flux}} = \mathbb{E}_{h_t} \left[D_{\mathrm{KL}} \left(p_\theta(\cdot \mid h_t) \| q_k^{\mathrm{flux}}(\cdot \mid h_t) \right) \right]. \tag{21}$$

Table 1: Results on the prompt optimization task with a Qwen3-VL-Instruct-8B teacher. The best and second-best student results are highlighted in bold and underlined, respectively.

Method	VBench ↑			VideoVerse ↑		
	VACE	CogVideoX	Average	VACE	CogVideoX	Average
Teacher	81.48	77.68	79.58	1105	1123	1114.0
Qwen3-VL-Instruct-2B						
Student	80.86	76.48	78.67	1092	1099	1095.5
+ OPD	<u>81.25</u> +0.39	<u>76.45</u> -0.03	<u>78.85</u> +0.18	1106 +14	<u>1097</u> -2	<u>1101.5</u> +6
+ OPCD	81.06 +0.20	76.07 -0.41	78.57 -0.10	<u>1107</u> +15	1078 -21	1092.5 -3
+ OEL†	79.97 -0.89	74.30 -2.18	77.14 -1.53	1106 +14	1022 -77	1064.0 -31.5
+ Flux-OPD	**81.50** +0.64	**77.10** +0.62	**79.30** +0.63	**1116** +24	**1106** +7	**1111.0** +15.5
Qwen3-VL-Instruct-4B						
Student	81.21	77.74	79.48	1107	1122	1114.5
+ OPD	81.02 -0.19	<u>77.75</u> +0.01	<u>79.39</u> -0.09	1081 -26	<u>1119</u> -3	1100.0 -14.5
+ OPCD	<u>81.09</u> -0.12	77.49 -0.25	79.29 -0.19	<u>1122</u> +15	1113 -9	<u>1117.5</u> +3
+ OEL†	79.77 -1.44	71.09 -6.65	75.43 -4.05	1103 -4	1041 -81	1072.0 -42.5
+ Flux-OPD	**81.91** +0.70	**77.95** +0.21	**79.93** +0.45	**1128** +21	**1122** +0	**1125.0** +10.5

5 Experiments

5.1 Implementation Details

Models We conduct experiments under different student-teacher settings: Qwen3-VL-Instruct (Yang et al., 2025a) with a 2B student and an 8B teacher, Qwen3-VL-Instruct with a 4B student and an 8B teacher, and Qwen3 with a 1.7B student and an 8B teacher. For each setting, the student is initialized from an SFT model trained for 5 epochs on offline teacher-distilled outputs, while the teacher uses the base model. All models operate in non-thinking mode.

Baselines We compare Flux-OPD with several OPD paradigms: (1) OPD (Agarwal et al., 2024), which distills teacher supervision on trajectories sampled from current student policy; (2) OPCD (Ye et al., 2026c), which conditions the teacher on privileged experience items as contexts; and (3) OEL (Ye et al., 2026b), which equips the teacher with evolving contexts for continuous student improvement. Moreover, on tasks with available rubrics, we compare RaR (Gunjal et al., 2025) with OPD paradigms by matching the number of student-generated completions, using the teacher as rubric judge (Appendix C.3). All baselines are implemented in a single-run training setting; we denote the resulting OEL adaptation as OEL†.

Tasks We focus on open-ended tasks with real-world applicability: prompt optimization for video generation and medical question answering, both involving complex preferences and requiring rich contextual information. Prompt optimization rewrites imperfect user prompts into generator-aligned prompts, requiring not only text-level teacher supervision but also contexts that convey video-level preferences of a specific generator. Medical question answering spans diverse user queries and clinical workflows, where contexts help capture domain knowledge and interaction preferences.

Training (1) Datasets. We use 10K SFT prompts from VPO (Cheng et al., 2025) for prompt optimization and RaR-Medicine (Gunjal et al., 2025) with 18K rubric-paired questions for medical question answering. (2) Context Extraction. Each context is extracted from 8 student rollouts with $M = 3$ random seeds per iteration. For prompt optimization, downstream videos generated from student responses serve as supplements, while paired rubrics serve as supplements for medical question answering. (3) Context Distillation. OPCD and OEL† use the same initial contexts as Flux-OPD. OEL† and Flux-OPD update contexts every 300 steps.

Evaluation (1) Prompt Optimization. We evaluate all methods with two downstream video generators from different model families, Wan2.1-VACE-1.3B and CogVideoX-2B (Jiang et al., 2025;

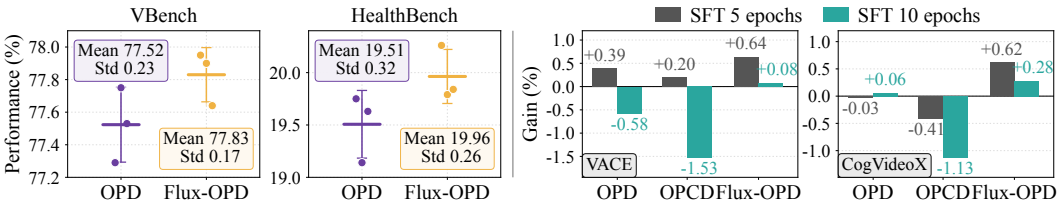

Figure 4: Left: Results across three independent training runs. Flux-OPD achieves higher mean scores than OPD. Right: Effect of student initialization. While stronger initialization reduces the gains from OPD and OPCD, Flux-OPD maintains performance improvements.

Yang et al., 2025b), on two video generation benchmarks: VBench and VideoVerse (Huang et al., 2024; Wang et al., 2025). (2) Medical Question Answering. We evaluate on HealthBench (Arora et al., 2025) to assess model safety and helpfulness in realistic medical scenarios.

5.2 Main Results

As shown in Tabs. 1 and 2, Flux-OPD achieves the best reported overall performance across the evaluated settings on two open-ended tasks. In medical question answering, where task preferences are primarily text-level, other OPD paradigms improve over the initial policy. While in prompt optimization, where video-level feedback provides generator-specific downstream signals, they may degrade the initial student policy, suggesting limited ability to effectively leverage such contexts. RaR also fails to improve under matched student-completion budget.

Moreover, we conduct three independent training runs for OPD and Flux-OPD. As shown in Fig. 4 (Left), Flux-OPD achieves higher mean scores than OPD on VBench (CogVideoX) and HealthBench, with comparable variation across runs.

Table 2: Results on the medical question answering task with a Qwen3 8B teacher. The best and second-best student results are highlighted in bold and underlined, respectively.

Method	HealthBench ↑	
	w/o Rubrics	w/ Rubrics
Teacher	37.38	-
Qwen3 1.7B		
Student	19.06	-
+ OPD	19.63 +0.57	-
+ OPCD	19.29 +0.23	19.53 +0.47
+ OEL†	<u>19.69</u> +0.63	<u>19.66</u> +0.60
+ RaR	-	16.91 -2.15
+ Flux-OPD	**20.20** +0.82	**20.26** +1.20

Moreover, we find that degradation is more common for the 4B student than for the 2B student. Since 5-epoch SFT may not fully absorb basic teacher knowledge, subsequent distillation gains may come from either general teacher supervision or contextual information. To isolate the effect of contextual supervision, we further evaluate OPD methods with a stronger 10-epoch SFT initialization. As shown in Fig. 4 (Right), stronger initialization generally reduces the gains of OPD and OPCD, with degradation particularly evident for OPCD. In contrast, Flux-OPD improves both initializations, indicating greater robustness. These results suggest that as students better absorb teacher knowledge, direct distribution matching provides diminishing benefits and may introduce harmful corrections, especially with context-conditioned teachers.

5.3 Analysis

We conduct additional analyses on VBench (4B student, 8B teacher) and HealthBench.

Effects of Evolving Contexts As shown in Tab. 3, full Flux-OPD achieves higher scores than V1, the variant w/o evolving contexts, demonstrating that contexts evolving with student performance improve the student more effectively under the same number of training steps.

Effects of Contextual Correction Contextual correction combines the context-free teacher with difference signals from context-conditioned teachers. As shown in Tab. 3, both OEL† and V2 have evolving contexts, while V2, equipped with contextual correction, achieves higher scores than OEL† under all three correction strengths. This demonstrates the effectiveness of using context-conditioned teachers as contextual difference signals rather than direct distillation targets.

Table 3: Ablation study of Flux-OPD. λ_k denotes the correction strength.

Variant	Evolving Contexts	Contextual Correction	Contextual Weighting	λ_k	VBench (VACE)	Health-Bench
V1	✗	✓	✓	Eq.(20)	81.29	19.63
OEL†	✓	✗	✗	-	79.77	19.66
V2	✓	✓	✗	$\equiv 0.9$	81.74	19.73
				$\equiv 0.7$	81.41	20.03
				$\equiv 0.5$	80.93	19.97
Full	✓	✓	✓	Eq.(20)	**81.91**	**20.26**

Table 4: Effect of the scaling factor β in Flux-OPD.

Method	β	VBench		Health-Bench
		VACE	CogVideoX	
Teacher	-	81.48	77.68	37.38
Student	-	81.21	77.74	19.06
Flux-OPD	8	81.27	**77.95**	19.27
	30	**81.91**	77.44	19.33
	50	81.49	77.54	**20.26**

Effects of Contextual Weighting Contextual weighting adjusts correction strengths based on the conflict term in Eq.(20). As shown in Tab. 3, full Flux-OPD achieves higher scores than all V2 variants with static strengths, showing that the weighting strategy is more effective than using a fixed strength. Moreover, it outperforms entropy- and JSD-based weighting under matched mean correction strength, supporting the use of the reverse-KL-derived conflict score (Appendix Sec. C.6).

Effect of the Scaling Factor The scaling factor β controls contextual correction based on conflicts among context-conditioned teachers. Prompt optimization involves additional variability from downstream video generation, whereas medical question answering relies on text-only interactions and exhibits more consistent contextual supervision (Appendix Fig. 13). Since more consistent contexts yield smaller conflict values and thus stronger corrections under the same β, a larger β can more strongly suppress contextual effects. Consistent with this observation, Tab. 4 shows that $\beta = 50$ gives the highest HealthBench score. Moreover, within prompt optimization, $\beta = 8$ performs better on CogVideoX by retaining stronger contextual correction. This is consistent with the observation that the initial student is already competitive with the context-free teacher.

5.4 Discussion

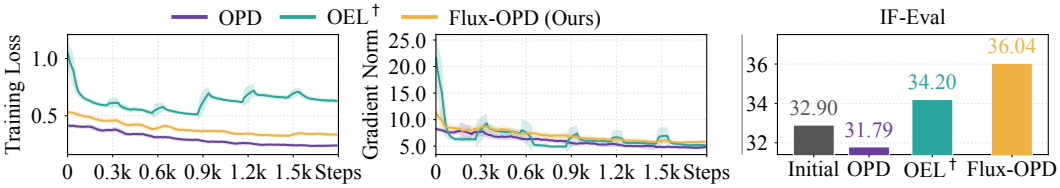

Figure 5: Left: Training stability. The training losses of OPD and Flux-OPD decrease steadily, whereas OEL† exhibits loss surges. Flux-OPD also maintains a stable gradient norm. Right: Out-of-distribution performance on IF-Eval. Flux-OPD achieves the highest prompt-level strict accuracy.

Stability As shown in Fig. 5 (Left), Flux-OPD exhibits stable training dynamics similar to OPD, while OEL† shows abrupt loss surges and larger gradient norm fluctuations around context updates.

Generalization We evaluate out-of-distribution instruction following on IF-Eval (Zhou et al., 2023) after training on medical question answering. As shown in Fig. 5 (Right), Flux-OPD achieves higher prompt-level strict accuracy than OPD, suggesting better cross-domain generalization.

6 Conclusion

We propose Flux-OPD, an OPD paradigm that uses evolving contexts as in-training supervision to incorporate contextual signals derived from task feedback. We characterize context-conditioned distillation through a reverse-KL decomposition and address distillation-target shifts and disagreement among context-conditioned teachers by anchoring distillation to the context-free teacher, injecting contextual difference signals, and weighting them according to teacher disagreement. These results highlight the potential to combine teacher supervision with evolving contexts.

AI USE STATEMENT

In this work, we used generative AI tools exclusively for grammatical error correction and language polishing. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide additional details to facilitate reproducibility in the appendix, including the implementation and training setup, dataset and evaluation details, hyperparameter configurations, and supplementary analyses. The derivations and additional explanations supporting our method are also provided in the appendix.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, et al. Vpo: Aligning text-to-video generation models with prompt optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15636–15645, 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*, 2023.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

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. *CoRR*, abs/2601.20802, 2026.

Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, et al. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18725–18735, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

John Schulman and Thinking Machines Lab. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250929. https://thinkingmachines.ai/blog/lora/.

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

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *CoRR*, abs/2209.15189, 2022. doi: 10.48550/ARXIV.2209.15189. URL https://doi.org/10.48550/arXiv.2209.15189.

Zeqing Wang, Xinyu Wei, Bairui Li, Zhen Guo, Jinrui Zhang, Hongyang Wei, Keze Wang, and Lei Zhang. Videoverse: Does your t2v generator have world model capability to synthesize videos? *arXiv preprint arXiv:2510.08398*, 2025.

Hongru Xiao, Jie Li, Zhirui Li, Bin Yang, Xiang Li, Sunzhu Li, and Jiale Han. Structuring human objectives: A survey of rubrics for evaluation, alignment, and agentic ai, June 2026. Available at SSRN: https://ssrn.com/abstract=6970459.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Tianzhu Ye, Li Dong, Guanheng Chen, He Zhu, Xun Wu, Shaohan Huang, and Furu Wei. Llm-as-a-coach: Experiential learning for non-verifiable tasks. *arXiv preprint arXiv:2607.18110*, 2026a.

Tianzhu Ye, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Online experiential learning for language models. *arXiv preprint arXiv:2603.16856*, 2026b.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026c.

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.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

APPENDIX

A PROMPT TEMPLATE

All prompts are intentionally concise and require **minimal human engineering**. They contain no human-crafted domain knowledge, allowing the models to rely on their own capabilities.

A.1 CONTEXT EXTRACTION PROMPT

Figs. 6 and 7 show the context extraction prompts used by OPCD, OEL†, and Flux-OPD for the prompt optimization and medical question answering tasks. These prompts extract new experience items beyond the existing ones.

A.2 CONTEXT DISTILLATION PROMPT

Figs. 8 and 9 show the in-context learning prompts used by context-conditioned teachers in OPCD, OEL†, and Flux-OPD for the prompt optimization and medical question answering tasks.

A.3 SYSTEM PROMPT

Fig. 10 shows the system prompt used for offline SFT data construction, student training, and evaluation in the prompt optimization task. We do not use an explicit system prompt for the medical question answering task.

B CONTEXT EXAMPLE

Figs. 11 and 12 show context examples from OPCD, OEL†, and Flux-OPD for the prompt optimization and medical question answering tasks.

C ADDITIONAL TRAINING DETAILS

C.1 HYPERPARAMETERS

General Hyperparameters For each training paradigm, we use a global batch size of 64 and a learning rate of 2×10^{-6}. SFT runs for 5 epochs, OPD-style methods run for 10 epochs, and RaR runs for 1.25 epochs with 8 completions per prompt. For computational efficiency, the implemented distillation loss retains the top-64 target logits and renormalizes them over the selected tokens. In Flux-OPD, this top-k truncation is applied only to the final distillation loss computation. The conflict term and the corresponding weighting signal are computed from full-vocabulary log-softmax distributions.

Hyperparameters of Flux-OPD We list the hyperparameters of Flux-OPD in Tab. 5. We use a smaller scaling factor for prompt optimization ($\beta = 30$) than for medical question answering ($\beta = 50$). As shown in Fig. 13, prompt optimization exhibits higher conflict values, partly due to the additional variability introduced by downstream video generation. Since larger conflict values already suppress contextual correction through Eq.(20), a smaller β avoids overly suppressing contextual effects.

C.2 CONTEXT

For each of the $M = 3$ random seeds, the teacher sequentially processes the same $n = 8$ trajectories. At each trajectory, it receives the experience items accumulated so far together with the user prompt, student response, and task-specific supplement (a generated video or medical rubrics), and is instructed to append task-general items not already covered. Each extraction call is capped at 256 new tokens; under this output specification, each context contains at most 2,048 generated tokens. Generation reaching this limit retains its prefix, and the accumulated context is not otherwise

Table 5: Additional hyperparameters of Flux-OPD.

Task	Student & Teacher	Downstream Model	β
Prompt Optimization	Qwen3-VL-Instruct 8B→2B	Wan2.1-VACE-1.3B	30
		CogVideoX-2B	30
	Qwen3-VL-Instruct 8B→4B	Wan2.1-VACE-1.3B	30
		CogVideoX-2B	8
Medical Question Answering	Qwen3 8B→1.7B	-	50

truncated. Deduplication is semantic and prompt-based, with no post-hoc lexical or embedding-based filtering. Contexts are rebuilt at each update rather than accumulated across iterations. Every distillation step uses all $M = 3$ contexts in the current pool, without subset sampling.

C.3 Compute and Budget Accounting

A teacher forward denotes one batched pass over the 64 student completions in an optimization step. OPD uses one context-free teacher forward per step, whereas Flux-OPD uses one context-free and $M = 3$ context-conditioned forwards, for four teacher forwards in total. Teacher logits are computed online and are not cached across optimization steps; only the extracted contexts are reused for 300 steps. At initialization and each context update, three contexts are extracted from eight student rollouts, requiring $3 \times 8 = 24$ teacher generation calls, each capped at 256 new tokens. The initial context pool is constructed once and shared across the context-based methods. For prompt optimization, the eight auxiliary videos are generated once per update and reused across the three extraction seeds. For medical question answering, the 10-epoch OPD-style runs generate one completion per prompt per epoch, while RaR generates 8 completions per prompt for 1.25 epochs. The executed runs contain $177{,}920$ and $177{,}472$ student completions, respectively, a 0.25% difference due to batch rounding. Thus, the comparison matches student-completion counts, not teacher-forward compute or total training FLOPs.

C.4 Effect of the Number of Contexts

Flux-OPD uses one context-free and M context-conditioned teacher forwards per step. As shown in Tab. 6, reducing M from 3 to 2 decreases the number of teacher forwards from 4 to 3, while changing VBench by only $+0.06$ on VACE and -0.20 on CogVideoX. Performance changes only marginally from $M = 2$ to $M = 3$, showing diminishing returns from additional context-conditioned teacher forwards. The larger degradation at $M = 1$ indicates the importance of using multiple contexts to estimate inter-context disagreement and apply contextual weighting.

Table 6: Effect of the number of contexts M with a Qwen3-VL-Instruct-8B teacher. Flux-OPD uses one context-free and M context-conditioned teacher forwards per optimization step.

M	Teacher Forwards per Step	VBench ↑	
		VACE (4B Student)	CogVideoX (2B Student)
1	2	77.84	75.62
2	3	**81.97**	<u>76.90</u>
3	4	<u>81.91</u>	**77.10**

C.5 Conflict Dynamics

We use the conflict metric to monitor distributional disagreement among context-conditioned teachers during training. For each response token position, the teacher is queried under multiple sampled contexts, producing multiple context-conditioned token distributions. These distributions are combined through a geometric mean in log-probability space. A larger conflict value indicates lower

distributional overlap and stronger disagreement among context-conditioned teachers. We report the average conflict over valid response tokens at each training step.

As shown in Fig. 13, medical question answering exhibits lower conflict values than prompt optimization, indicating more consistent contextual supervision. The higher conflict values in prompt optimization may result from the additional variability introduced by downstream video generation, which can reduce the consistency of the resulting contexts.

C.6 Alternative Weighting Criteria

To compare our reverse-KL-derived conflict score with generic weighting criteria, we consider two alternatives. Let $\{p_k^{(m)}\}_{m=1}^M$ denote the context-conditioned teacher distributions at token position k. We define the entropy and Jensen-Shannon divergence (JSD) scores as

$$s_k^{\mathrm{Ent}} = \frac{1}{M}\sum_{m=1}^{M} H\left(p_k^{(m)}\right), \tag{22}$$

and

$$s_k^{\mathrm{JSD}} = H(\bar{p}_k) - \frac{1}{M}\sum_{m=1}^{M} H\left(p_k^{(m)}\right), \qquad \bar{p}_k = \frac{1}{M}\sum_{m=1}^{M} p_k^{(m)}. \tag{23}$$

where $H(p) = -\sum_v p(v)\log p(v)$. The entropy score measures the uncertainty of context-conditioned teachers, while the JSD score measures their distributional disagreement. Each score is mapped to a contextual weight by

$$\lambda_k^r = \exp(-\alpha_r s_k^r), \qquad r \in \{\mathrm{Ent}, \mathrm{JSD}\}. \tag{24}$$

Because these scores have different numerical scales, we avoid downstream tuning and choose α_r for each training micro-batch such that the mean weight over valid response tokens matches that produced by Flux-OPD:

$$\frac{1}{|\mathcal{B}|}\sum_{k\in\mathcal{B}} \exp(-\alpha_r s_k^r) = \frac{1}{|\mathcal{B}|}\sum_{k\in\mathcal{B}} \exp(-\beta\delta_k), \tag{25}$$

where $\mathcal{B}$ denotes the set of valid response-token positions in the micro-batch. All other training and evaluation settings are unchanged, so the comparison keeps the mean contextual weight matched while varying the weighting criterion. As shown in Tab. 7, the reverse-KL-derived conflict score outperforms entropy and JSD weighting by 0.55 and 0.15 points, respectively. This result suggests that the reverse-KL-derived conflict score is more effective than these generic uncertainty- and disagreement-based alternatives.

Table 7: Comparison of token-level weighting criteria for Qwen3-VL-Instruct 8B→2B, evaluated on VBench with CogVideoX-2B. For each baseline, the mean contextual weight is matched to that of Flux-OPD within every training micro-batch.

Weighting Criterion	**VBench ↑**
Entropy	76.55
JSD	<u>76.95</u>
Reverse-KL Conflict (Ours)	**77.10**

D Additional Evaluation Details

For the prompt optimization task, we evaluate the 10 VBench dimensions (Huang et al., 2024) reported in prior works (Cheng et al., 2025; Ji et al., 2025). The quality dimensions include Subject Consistency, Background Consistency, Motion Smoothness, Dynamic Degree, Aesthetic Quality, and Image Quality, while the semantic dimensions include Multiple Objects, Human Action, Scene, and Appearance Style. These dimensions cover 523 prompts in total. We evaluate all dimensions of the full VideoVerse (Wang et al., 2025), covering 300 prompts. VBench reports scores based on automatic metrics, whereas VideoVerse uses Qwen2.5-VL-32B-Instruct for scoring following the official protocol.

Context Extraction Prompt - Prompt Optimization

You are an AI language model that continuously refines its internal experience.

Previous experience:
{previous_experience}

Here is the latest interaction history (the user's original prompt, your rewritten prompt, and the sampled video generated from the rewritten prompt):
{latest_experience}

Your task:
Based on the latest interaction history and the previous experience, generate additional experience for future prompt rewriting.
You should analyze what makes a rewritten prompt better or worse, considering both the rewritten prompt itself and the quality of the sampled video generated from it, infer the general principles behind effective rewriting, and summarize these insights into 1-2 concise, high-level, and widely applicable experience items that help improve future rewriting performance.

Rules:
- The experience you generate MUST be formatted strictly as a markdown list where each item starts with "- EXPERIENCE ITEM:".
- The experience you generate will be directly appended to the previous experience. Do not repeat the previous experience. Make sure the newly generated experience is different from the previous experience.
- Your generated experience should capture general rewriting strategies, principles, or rules, rather than details specific to the current example.
- Use the sampled video as evidence of whether the rewritten prompt helped or hurt generation quality, faithfulness, clarity, or controllability, but summarize only the general lesson for future prompt rewriting.
- The change should introduce a high-level, broadly useful insight about prompt rewriting, such as preserving user intent, improving clarity, controlling expansion, avoiding unsupported additions, or handling ambiguity conservatively. The updated experience must remain concise, structured, and meaningful.
- If the new insight conflicts with any previous experience item, you may describe the conflict and provide a resolution in the new item.

Output the final result in exactly this format:
- EXPERIENCE ITEM: ...
- EXPERIENCE ITEM: ...

Figure 6: Prompt for context extraction in the prompt optimization task.

Context Extraction Prompt - Medical Question Answering

You are an AI language model that continuously refines its internal experience.

Previous experience:
{previous_experience}

Here is the latest interaction history (the user's prompt, your response, and the rubrics):
{latest_experience}

Your task:
Based on the latest interaction history and the previous experience, generate additional experience for future learning.
You should analyze what makes your response better or worse, considering both the response itself and the rubrics, infer the general principles behind effective conversation, and summarize these insights into 1-2 concise, high-level, and widely applicable experience items that help improve future performance.

Rules:
- The experience you generate MUST be formatted strictly as a markdown list where each item starts with "- EXPERIENCE ITEM:".
- The experience you generate will be directly appended to the previous experience. Do not repeat the previous experience. Make sure the newly generated experience is different from the previous experience.
- Your generated experience should capture general response strategies, principles, or rules, rather than details specific to the current example.
- Use the rubrics as evidence of whether the response helped or hurt, but summarize only the general lesson for future learning.
- The change should introduce a high-level, broadly useful insight. The updated experience must remain concise, structured, and meaningful.
- If the new insight conflicts with any previous experience item, you may describe the conflict and provide a resolution in the new item.

Output the final result in exactly this format:
- EXPERIENCE ITEM: ...
- EXPERIENCE ITEM: ...

Figure 7: Prompt for context extraction in the medical question answering task.

Context Distillation Prompt - Prompt Optimization

Rewrite the user's prompt into one clear, concrete prompt for video generation. Output only the rewritten prompt.

Here are experiential knowledge items for reference:
{CONTEXT}

Original prompt:

Figure 8: Prompt for context distillation in the prompt optimization task.

Context Distillation Prompt - Medical Question Answering

You are a helpful assistant.

Here are experiential knowledge items for reference:
{CONTEXT}

User prompt:

Figure 9: Prompt for context distillation in the medical question answering task.

System Prompt - Prompt Optimization

Rewrite the user's prompt into one clear, concrete prompt for video generation. Output only the rewritten prompt.

Original prompt:

Figure 10: System prompt for the prompt optimization task.

Context - Prompt Optimization

- EXPERIENCE ITEM: Prioritize preserving the core narrative intent while enhancing specificity to guide generation toward desired emotional or relational dynamics, as vague terms like “amazed” can lead to generic or mismatched visual interpretations.
- EXPERIENCE ITEM: Avoid introducing unsupported or unrequested elements (e.g., characters, actions, or settings not implied in the original prompt) that may misdirect the model or dilute the user’s focus, even if they seem plausible or complementary.
- EXPERIENCE ITEM: Prioritize structural clarity in rewritten prompts by explicitly separating visual elements (e.g., lighting, motion, texture) from narrative or emotional intent, to prevent ambiguous blending that may misdirect generation toward unintended aesthetics or compositions.
- EXPERIENCE ITEM: When rewriting prompts, prioritize explicit control over ambiguous terms by replacing vague descriptors (e.g., “lively,” “heartily”) with concrete, actionable visual or behavioral cues (e.g., “clapping,” “laughing with open mouths”) to ensure generation aligns precisely with the user’s intent.
- EXPERIENCE ITEM: When the sampled output diverges from the prompt’s core subject (e.g., showing a different character or setting), prioritize explicitly naming the intended subject and setting in the rewritten prompt to prevent misalignment, even if it requires adding minimal contextual anchors like “in a mythological battle scene” or “as Perseus.”
- EXPERIENCE ITEM: .....

Figure 11: Context example for the prompt optimization task.

Context - Medical Question Answering

- EXPERIENCE ITEM: Differentiate between respiratory infections based on the presence of blood in sputum, acute onset, and systemic toxicity, as these can point to conditions like pneumonic plague rather than more chronic or less severe infections like tuberculosis or Legionella.
- EXPERIENCE ITEM: When calculating diagnostic test metrics like specificity, it is crucial to distinguish between true negatives and false positives, as specificity reflects the test's ability to correctly identify those without the disease.
- EXPERIENCE ITEM: When interpreting clinical scenarios involving complex pathophysiology, it is essential to focus on the core mechanisms of disease progression, such as viral replication steps, rather than getting distracted by secondary symptoms or diagnostic findings, as this helps in identifying the correct pathogen and its associated enzymes.
- EXPERIENCE ITEM: When selecting medications for complex patient profiles, prioritize **safety profiles** and **drug interactions** to avoid **compromising therapeutic outcomes** or **increasing the risk of adverse events**, especially in patients with **multiple chronic conditions**.
- EXPERIENCE ITEM: When evaluating patients with complex hematologic findings, it is essential to **correlate clinical symptoms with potential drug exposures** and **consider the pharmacologic properties of medications** in the context of the patient’s medical history, as this helps in identifying toxic or idiosyncratic drug reactions that may be the underlying cause of the abnormal findings.
- EXPERIENCE ITEM: .....

Figure 12: Context example for the medical question answering task.

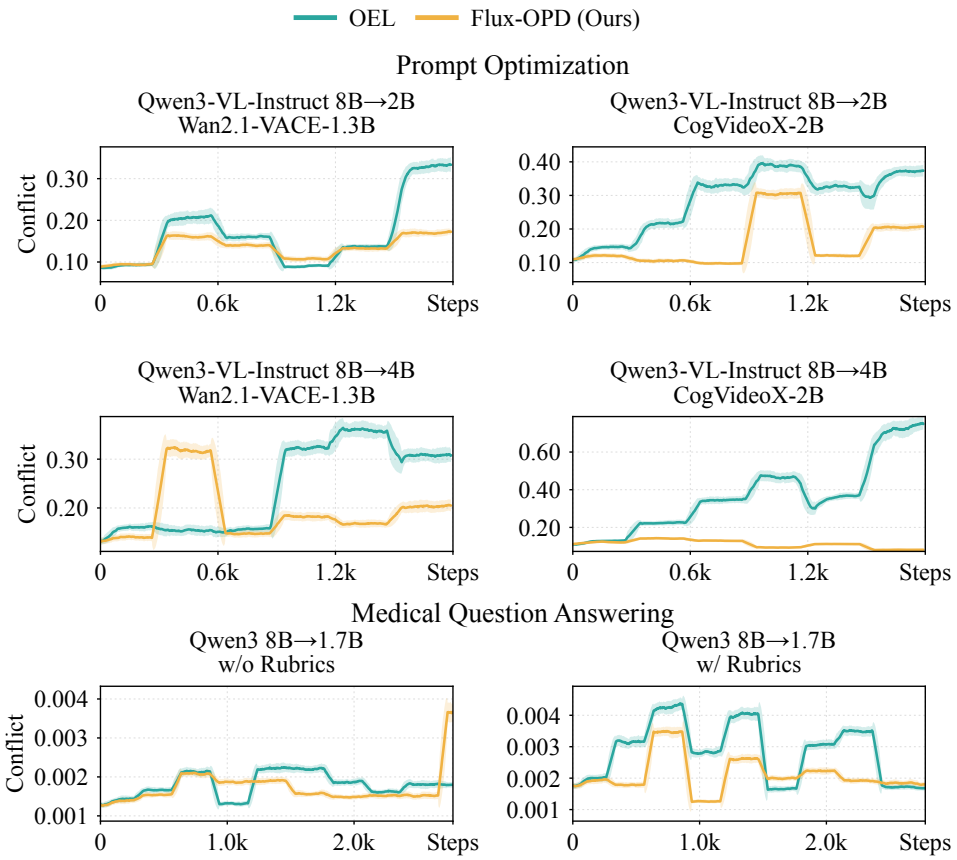

Figure 13: Conflict dynamics. The prompt optimization settings exhibit higher conflict values, whereas medical question answering maintains lower conflict, potentially because video sampling introduces greater variability and reduces context consistency.