CCC: Enhancing Video Generation via Structured MLLM Feedback

Jing Gu [1] **Ashwin Nagarajan** [1] **Tejas Polu** [1] **Kaizhi Zheng** [1] **Ruijian Zha** [2] **Jie Yang** [3] **Xin Eric Wang** [1 †]

Abstract

Video generation from natural-language prompts has made impressive strides, but current systems frequently misalign outputs with their input descriptions, dropping critical details, and hallucinating unintended content. Existing approaches to improving video quality typically rely on heavyweight post-editing models, which may introduce new artifacts, or costly fine-tuning of the generator backbone, limiting scalability and accessibility. While multimodal large language models (MLLMs) have demonstrated strong capabilities in diagnosing visual-text misalignment, their use has largely focused on image-level improvement rather than video. Therefore, we introduce `Critique Coach Calibration` (`CCC`), a training-free, test-time prompt-adaptation framework that closes the loop between generation and evaluation. In each iteration, an off-the-shelf MLLM produces a structured critique of a generated video, highlighting misaligned semantics, subject drift, and missing objects, and then reformulates the input prompt based on its own feedback. By repeating this critique–coach cycle, `Critique Coach Calibration` drives steady improvements in video quality without modifying the generator or relying on external editing modules. Empirical results on diverse video scenarios demonstrate that our approach consistently enhances semantic alignment and visual quality.

Figure 1: **Two examples of `Critique Coach Calibration` (`CCC`)'s impact:** *Top:* The initial generation omits the children; after `CCC`, they appear. *Bottom:* The first clip lacks the javelin; `CCC` restores it.

1. Introduction

Recent text-to-video models can produce compelling clips from language prompts, yet they often omit key details, hallucinate extraneous content, or break temporal consistency (Figure 1). Existing remedies either require expensive architectural modifications or introduce artifacts via localized editing. Multimodal LLMs, however, excel at diagnosing visual–textual mismatches, suggesting a path to automated, model-agnostic refinement without retraining.

We introduce `Critique Coach Calibration` (`CCC`), a training-free loop that uses a frozen MLLM as both critic and coach. In each iteration, the critic issues a structured report (e.g. flagging missing objects, semantic drift, and quality glitches) and the coach rewrites the prompt to address these issues. By repeating generation, diagnosis, and prompt refinement, `CCC` steadily improves semantic alignment and visual fidelity without any model-specific tuning.

Our experiments show that two iterations of `CCC` suffice to outperform one-shot baselines across diverse video scenarios, delivering significant gains in alignment fidelity and overall quality.

2. Related Work

Most text-to-video repair methods attack one failure mode at a time. VideoRepair, for instance, locally re-diffuses misaligned regions but cannot rewrite prompts or fix temporal drift (Lee et al., 2025). Our `Critique Coach Calibration` (`CCC`) loop instead pairs a frozen text-to-video model

[1]University of California, Santa Cruz [2]Columbia University [3]Cybever. Correspondence to: Jing Gu <jgu110@ucsc.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

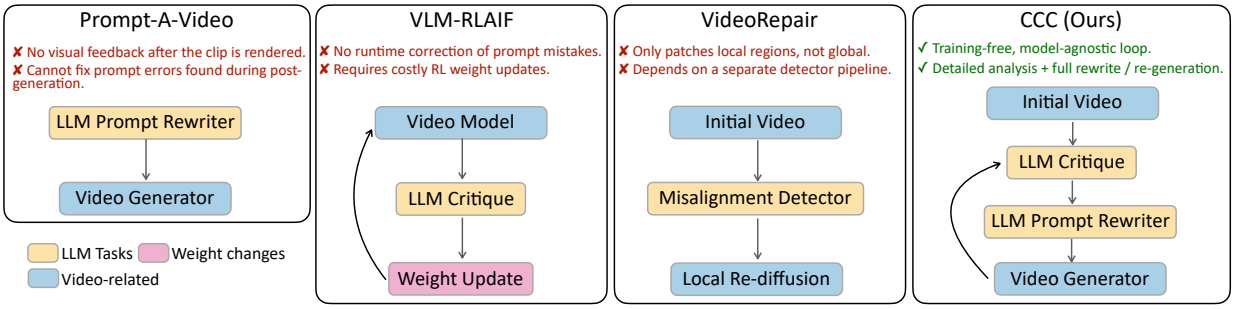

Figure 2: **Training-free, model-agnostic refinement loop.** Prompt-A-Video, VLM-RLAIF, and VideoRepair either lack visual feedback, require RL weight updates, or patch only local regions. Our loop uses an LLM critique to rewrite the *entire* prompt, then fully re-generates the clip with no parameter tuning.

with a single multimodal LLM that first *critiques* a draft clip and then *coaches* a prompt rewrite, regenerating the entire video. As Figure 2 illustrates, CCC unifies the three previously disjoint stages of the literature, pre-generation prompt search (Prompt-A-Video, VPO), training-time alignment (VLM-RLAIF, DPO-Video), and post-generation patching (VideoRepair, CSR) (Ji et al., 2024; Cheng et al., 2025; Ahn et al., 2024; Zhang et al., 2025; Zhou et al., 2024). By iterating this critique $\rightarrow$ rewrite loop and accepting updates only when self-consistency holds, CCC delivers global improvements without any weight tuning.

The same LLM that rewrites prompts also serves as a zero-shot evaluator, subsuming the role of prior video judges such as AIGV-Assessor, LMM-VQA, UVE, Evaluation Agent, VideoAutoArena, and LLaVA-Critic (Wang et al., 2024; Ge et al., 2024; Liu et al., 2025; Zhang et al., 2024; Luo et al., 2025; Xiong et al., 2024). We extend self-consistency decoding (Wang et al., 2023) and LLM-uncertainty ideas (Xie et al., 2025) to the multimodal setting: each prompt–video pair is queried K times, yielding an overall agreement score and a fine-grained Content Agreement Score that jointly gate further refinements. In this way, CCC closes the evaluation $\rightarrow$ repair loop using a single, black-box LLM while remaining model-agnostic to the underlying video generator.

3. Critique Coach Calibration

Critique Coach Calibration (CCC) wraps a frozen text-to-video model in a self-correcting loop (Fig. 3). At round t, prompt $\mathcal{P}^{(t)}$ produces $\mathcal{V}^{(t)} = G_\theta(\mathcal{P}^{(t)})$. The CRITIC assigns a coarse score y and issue list $\mathcal{C}^{(t)}$, and the COACH consumes $(\mathcal{P}^{(t)}, \mathcal{C}^{(t)})$ to output a corrective rewrite $\mathcal{R}^{(t)}$. We set

$$\mathcal{P}^{(t+1)} = \mathcal{P}^{(t)} \oplus \mathcal{R}^{(t)}$$

and repeat for T rounds.

3.1. Reliability of the CRITIC

We sample F frames per video and query the MLLM K times for stability. The critic yields scores $\{y_k\}$ whose mode defines agreement:

$$\gamma = \frac{\left|\{k : y_k = \text{mode}(y)\}\right|}{K} \tag{1}$$

For issue sets $\mathcal{I}_k$, define

$$A = \left|\bigcap_{k=1}^{K} \mathcal{I}_k\right|, \quad B = \left|\bigcup_{k=1}^{K} \mathcal{I}_k\right| - A, \tag{2}$$

$$C = \sum_{k=1}^{K} |\mathcal{I}_k| - A - B. \tag{3}$$

The Content Agreement Score is

$$\Psi_{\text{CAS}} = \frac{A + 0.5\,B}{A + B + C}. \tag{4}$$

Full loop. Algorithm 1 summarizes the complete CCC procedure.

4. Experimental Results

4.1. Setup

We evaluate CCC on two text-to-video backbones (LTX-Video, Wanx-2.1) over three calibration rounds ($T = 3$). Each round incurs one generation pass (≈15 seconds per clip for LTX, 120 seconds for Wanx). Human judgments on Amazon Mechanical Turk assess (i) semantic alignment and (ii) visual quality; we also report automatic alignment scores via EvalCrafter (Liu et al., 2024).

4.2. Main Results

Table 1 shows that annotators prefer using the "negative strategy" over the baseline outputs as seen by the improvements

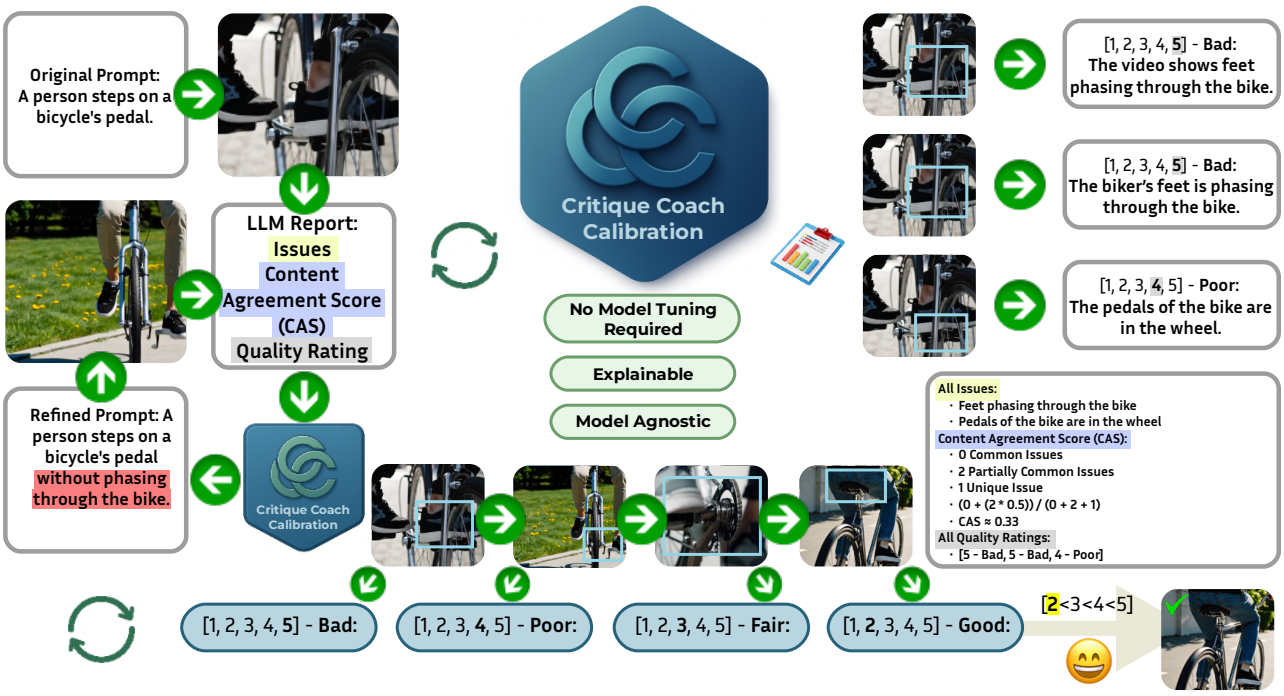

Figure 3: **Overview of `Critique Coach Calibration` (CCC):**. MLLM first issues K self-consistent critiques of a generated clip, aggregates recurring failure modes into severity tiers (e.g., *Major*, *Minor*), and then distills the dominant findings into calibrated guidance that steers the next generation cycle. The bounding boxes are for visualization only, not generated in our pipeline.

Table 1: **Human evaluation results.** Percent of raters preferring our model over baselines. **SA** = semantic alignment, **VQ** = video quality, **-N** = without negative strategy. Numbers in parentheses denote change (Δ) relative to the "-N" baseline.

	LTX-Video			Wanx-2.1		
	Origin	OPT2I	LLM-P	Origin	OPT2I	LLM-P
SA-N	61.4%	55.7%	56.6%	59.7%	62.9%	63.4%
VQ-N	51.7%	65.7%	55.4%	57.1%	58.0%	55.4%
SA	62.8% (Δ+1.4%)	57.1% (Δ+1.4%)	62.9% (Δ+6.3%)	60.3% (Δ+0.6%)	56.9% (Δ–6.0%)	63.1% (Δ–0.3%)
VQ	54.5% (Δ+2.8%)	64.0% (Δ–1.7%)	57.7% (Δ+2.3%)	59.7% (Δ+2.6%)	58.5% (Δ+0.5%)	58.9% (Δ+3.5%)

Table 2: **Automatic evaluation results.** `Critique Coach Calibration` boosts the performance of LTX-Video and Wanx-2.1 on semantic alignment.

Method	**LTX-Video**			**Wanx-2.1**		
	Origin	OPT2I	LLM-P	Origin	OPT2I	LLM-P
OPT2I	43.0	45.1	46.1	46.8	48.0	46.5
LLM-P	42.5	46.1	45.2	43.1	47.2	47.4
Origin	40.2	44.1	45.1	45.7	47.3	46.7
CCC (Ours)	**46.1**	**47.3**	**48.1**	**49.2**	**49.0**	**48.2**

in percentages. Table 2 shows that our method performs better than others in semantic alignment. Qualitatively (Fig. 4), it can be seen that `Critique Coach Calibration` corrects red-boxed semantic errors and red-circled visual artifacts that persist in other baselines.

4.3. Effectiveness of the Negative Strategy

To probe how framing influences critique quality, we compare a "negative" prompt that warns the MLLM of potential errors (e.g., "There could be issues...") against a neutral, direct prompt. In a controlled human evaluation, the negative prompt elicits richer, more accurate issue reports—surfacing omissions that the neutral prompt overlooks.

Quantitatively, Table 1 shows that adding this fault-seeking clause improves both semantic alignment and perceptual quality across our two backbones. This boost underscores the value of explicitly priming the critic to search for errors before generating its structured feedback.

4.4. Ablation Study

We conduct three ablations to isolate the impact of key components in our pipeline. In the *No Self-Consistency* variant, we skip the repeated sampling and content agreement analysis, using only the first MLLM response for prompt refinement. In the *No Negative Strategy* variant, we omit the initial error-warning phrasing and rely on a neutral direct

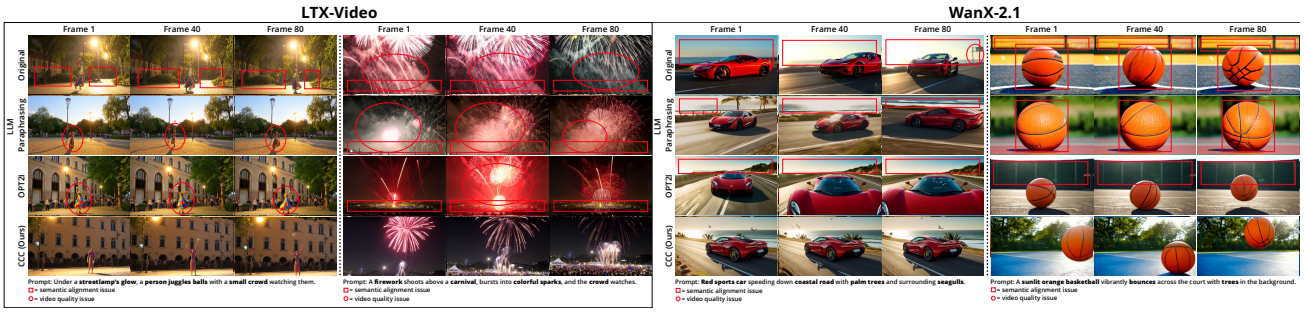

Figure 4: **Qualitative Comparison with Baselines.** We showcase four illustrative cases—two produced with LTX-Video and two with Wanx-2.1. While the baselines strive to enhance video quality, CCC yields markedly superior generations.

Algorithm 1 CCC Loop (cf. Figure 3)

Input: Initial prompt $\mathcal{P}^{(0)}$, generator G_θ, critic/coach LLM ϕ, iterations T, repetitions K, samples per round M
for $t = 0$ to $T-1$ **do**
  Sample candidate videos $\{V_i^{(t)}\}_{i=1}^{M} \leftarrow G_\theta(\mathcal{P}^{(t)})$ ▷ Generate initial candidate videos
  for $i = 1$ to M **do**
    Query $\{(y_{i,k}, \mathcal{C}_{i,k})\}_{k=1}^{K}$ from ϕ ▷ Obtain K critic /score pairs
    Compute γ_i via Eq. (1) and $\Psi_{\text{CAS},i}$ via Eq. (4) ▷ Compute confidence and CAS metrics
    Select canonical critique $\mathcal{C}_i^{(t)}$ with highest $\Psi_{\text{CAS},i}$ ▷ Keep critique with highest CAS
  end for
  $i^* \leftarrow \arg\max_i \Psi_{\text{CAS},i}$ ▷ Choose clip with best CAS
  $\mathcal{C}^{(t)} \leftarrow \mathcal{C}_{i^*}^{(t)}$ ▷ Use its critique as canonical
  $\mathcal{R}^{(t)} \leftarrow f_\phi\big(\mathcal{P}^{(t)}, \mathcal{C}^{(t)}\big)$ ▷ Produce rewrite via critic
  $\mathcal{P}' \leftarrow \mathcal{P}^{(t)} \oplus \mathcal{R}^{(t)}$ ▷ Augment prompt with rewrite
  Sample refined videos $\{V_j'^{(t)}\}_{j=1}^{M} \leftarrow G_\theta(\mathcal{P}')$ ▷ Generate refined videos
  for $j = 1$ to M **do**
    Rate each refined clip yielding y_j' ▷ Critic rates each refined clip
  end for
  $j^* \leftarrow \arg\min_j y_j'$ ▷ Select best refined clip
  $V^{(t)} \leftarrow V_{j^*}'^{(t)}$ ▷ Save chosen clip
  $\mathcal{P}^{(t+1)} \leftarrow \mathcal{P}'$ ▷ Set prompt for next iteration
end for
Output: Final prompt $\mathcal{P}^{(T)}$ and chosen videos $\{V^{(t)}\}_{t=0}^{T-1}$

prompt for evaluation. In the *No Iterative Loop* variant, we perform only one round of evaluation and refinement rather than iterating the feedback cycle. This comparison reveals how much each component—self-consistency, negative strategy, and iterative looping—contributes to the final video quality.

4.5. Generalization to Other Models

Applying the same $M = 3$, $F = 8$ loop to additional models yields consistent gains over raw generation, demonstrating that CCC's multi-round critique–rewrite pipeline is broadly effective across text-to-video models.

Table 3: Ablation results: effect of removing key components in the prompt optimization process. It shows win rate of full method of CCC over the ablation.

Ablation Variant	Alignment with Human
No Self-Consistency	55.0
No Iteration Loop	61.2
No Both	67.1

5. Conclusion

We propose Critique Coach Calibration (CCC), a lightweight loop that lets a frozen text-to-video model learn from an off-the-shelf multimodal LLM: the LLM first critiques a generated clip, then rewrites the prompt, and the cycle repeats. Across two backbones, CCC lifts human-rated semantic alignment and visual quality significantly, without retraining or heavy post-editing. The method is transparent, cheap to deploy, and ablation studies show each design choice matters. Limitations include missed transient errors and dependence on proprietary critics; future work can add motion-aware scoring and open-weight evaluators. We hope CCC encourages broader closed-loop generation systems that let evaluators directly coach generators.

Impact Statement

This paper introduces Critique Coach Calibration, an automated, model-agnostic loop that iteratively critiques and rewrites prompts for text-to-video models, improving general issues found in the videos. The approach can benefit educational media, bias-controlled synthetic datasets for vision tasks.

References

Ahn, D., Choi, Y., Yu, Y., Kang, D., and Choi, J. Tuning large multimodal models for videos using reinforcement learning from AI feedback. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 923–940, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.52. URL https://aclanthology.org/2024.acl-long.52/.

Cheng, J., Lyu, R., Gu, X., Liu, X., Xu, J., Lu, Y., Teng, J., Yang, Z., Dong, Y., Tang, J., Wang, H., and Huang, M. Vpo: Aligning text-to-video generation models with prompt optimization. *CoRR*, abs/2503.20491, March 2025. URL https://doi.org/10.48550/arXiv.2503.20491.

Ge, Q., Sun, W., Zhang, Y., Li, Y., Ji, Z., Sun, F., Jui, S., Min, X., and Zhai, G. Lmm-vqa: Advancing video quality assessment with large multimodal models, 2024. URL https://arxiv.org/abs/2408.14008.

Ji, Y., Zhang, J., Wu, J., Zhang, S., Chen, S., GE, C., Sun, P., Chen, W., Shao, W., Xiao, X., Huang, W., and Luo, P. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm, 2024. URL https://arxiv.org/abs/2412.15156.

Lee, D., Yoon, J., Cho, J., and Bansal, M. Videorepair: Improving text-to-video generation via misalignment evaluation and localized refinement, 2025. URL https://arxiv.org/abs/2411.15115.

Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., and Shan, Y. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024.

Liu, Y., Zhu, R., Ren, S., Wang, J., Guo, H., Sun, X., and Jiang, L. Uve: Are mllms unified evaluators for ai-generated videos?, 2025. URL https://arxiv.org/abs/2503.09949.

Luo, Z., Wu, H., Li, D., Ma, J., Kankanhalli, M., and Li, J. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation, 2025. URL https://arxiv.org/abs/2411.13281.

Wang, J., Duan, H., Zhai, G., Wang, J., and Min, X. Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm, 2024. URL https://arxiv.org/abs/2411.17221.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Xie, Q., Li, Q., Yu, Z., Zhang, Y., Zhang, Y., and Yang, L. An empirical analysis of uncertainty in large language model evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=J4xLuCt2kg.

Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., and Li, C. LLaVA-critic: Learning to evaluate multimodal models, 2024. URL https://openreview.net/forum?id=L4nH3j7L94.

Zhang, F., Tian, S., Huang, Z., Qiao, Y., and Liu, Z. Evaluation agent: Efficient and promptable evaluation framework for visual generative models, 2024. URL https://arxiv.org/abs/2412.09645.

Zhang, R., Gui, L., Sun, Z., Feng, Y., Xu, K., Zhang, Y., Fu, D., Li, C., Hauptmann, A. G., Bisk, Y., and Yang, Y. Direct preference optimization of video large multimodal models from language model reward. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 694–717, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.30/.

Zhou, Y., Fan, Z., Cheng, D., Yang, S., Chen, Z., Cui, C., Wang, X., Li, Y., Zhang, L., and Yao, H. Calibrated self-rewarding vision language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 51503–51531. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5c20c00504e0c049ec2370d0cceaf3c4-Paper-Conference.pdf.

A. MLLM Prompts

We designed three closely related prompt templates to elicit both an error diagnosis and a corrective rewrite from the critic LLM. In each case, the prompt begins by identifying the input as an AI-generated video and including the original generation prompt (`current_prompt`). The first template reads:

> `This is a video generated by a video generation model. The video is generated by prompt: '<current_prompt>'. Please give a simple description of the issue in this video, and also give a new prompt to generate a new video to avoid the issue. The new prompt should not miss any details in the original prompt. Please return the description in JSON format. The JSON format should be like this: {'issue': 'description', 'new_prompt': 'new_prompt'}.`

We then extended this to explicitly acknowledge potential generation artifacts by adding the clause "Since it is a generated video, so there could be some issues," yielding the second template:

> `This is a video generated by a video generation model. The video is generated by prompt: '<current_prompt>'. Since it is a generated video, so there could be some issues. Please give a simple description of the issue in this video, and also give a new prompt to generate a new video to avoid the issue. Do not miss any details in the original prompt. Please return the description in JSON format. The JSON format should be like this: {'issue': 'description', 'new_prompt': 'new_prompt'}.`

Finally, we introduced a quality rating field to capture coarse perceptual quality alongside the issue and rewrite, resulting in the third template:

> `This is a video generated by a video generation model. The video is generated by prompt: '<current_prompt>'. Since it is a generated video, so there could be some issues. Please give a simple description of the issue in this video, and also give a new prompt to generate a new video to avoid the issue. Do not miss any details in the original prompt. Please return the description in JSON format. The JSON format should be like this: {'quality': '1. Excellent, 2. Good, 3. Fair, 4. Poor, 5. Bad', 'issue': 'description', 'new_prompt': 'new_prompt'}.`

For consistency with human evaluation scales, we interpret the quality field according to the ordered ratings "1. Excellent, 2. Good, 3. Fair, 4. Poor, 5. Bad." All three templates were applied to every video in our corpus to compare how explicit mention of potential issues and inclusion of a quality rating affect the critic's diagnostic output and the subsequent prompt refinement.

B. Implementation Details

Our entire pipeline is organized into three conceptual stages:

Frame Sampling and Encoding. Each input video is first sampled to extract a fixed number of evenly spaced key frames. These frames are prepared for the critic LLM by converting them into a format suitable for multimodal input.

LLM Critique Interface. We wrap all interactions with the frozen multimodal LLM (GPT-4o) in a lightweight interface that: (1) fills one of the predefined prompt templates with the current generation prompt and sampled frames; (2) issues multiple independent calls under identical inputs to obtain a set of critiques; and (3) parses the returned JSON or structured text into categorical labels and issue lists for downstream processing.

`Critique Coach Calibration` Loop. The core loop (Alg. 1) alternates between *critique* and *rewrite* phases. In the critique phase, we aggregate repeated LLM outputs to compute both a confidence score (self-consistency) and a content agreement score (variance analysis). In the rewrite phase, a second prompt template fuses the canonical critique back with the original prompt to produce a refined description. This two-stage cycle is repeated for a fixed number of iterations, yielding progressively improved prompts.

All configuration—number of sampled frames, self-consistency repetitions, iteration count, and choice of prompt templates—is exposed as a simple set of parameters. This modular design allows straightforward swapping of alternative prompt styles or LLM backends without touching the core orchestration logic.

C. Baseline Implementation Details

We benchmark `CCC` against three *training-free* prompt-refinement baselines that keep **every** non-language-model hyper-parameter fixed—namely the frozen generator G_θ, the clip budget (M), the key-frame count (F), and the GPT judge configuration described in Section 4.1. *Origin* forwards the raw prompt $P^{(0)}$ to G_θ and selects the best-rated clip, providing a no-feedback reference. *LLM-P* performs a single text-only paraphrase: GPT-4o rewrites $P^{(0)}$ *without* seeing any video, after which M new clips are generated and scored. *OPT2I* executes one round of visual feedback: GPT-4o first paraphrases $P^{(0)}$, inspects a draft clip, issues a second rewrite $\hat{P}$, and

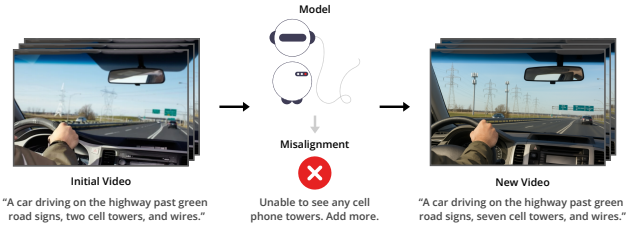

Figure 5: **Failure Case 1: Undetected objects trigger escalating hallucination.** When the critic fails to see the small towers in the initial clip, it requests to "add more," causing the refined prompt to overshoot from two to seven towers in the next iteration.

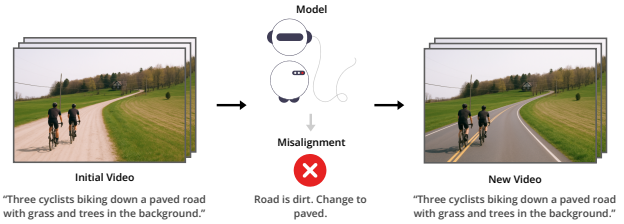

Figure 6: **Failure Case 2: A mislabeled surface type drives cascading feedback errors.** The ground-truth scene contains a dirt path, yet the prompt inaccurately calls it paved. Trusting this label, the critic instructs the generator to add pavement, so the refinement step paves the scene and the video drifts further from reality.

we regenerate M clips from $G_\theta(\hat{P})$. Because these variants differ only in the amount of critique and visual context (none, text-only, or one visual round), the gaps to CCC reported in Figure 4 isolate the value of our multi-round critique-and-rewrite loop.

D. Failure Cases

Even though our evaluation-refinement loop is very successful, there are a few problems that deserve a deeper study.

Low-quality input can nudge the loop off track (Fig. 5). Compression artifacts, motion blur or noise may obscure small details (e.g. distant towers or wires) so the critic misreports missing elements and the prompt rewriter "chases" phantom objects in subsequent iterations. A lightweight prefilter (e.g. denoising or super-resolution) or a confidence flag for low-quality frames would help prevent this mild but repetitive drift.

Prompt–video mismatches amplify over rounds (Fig. 6). If the initial prompt mislabels a scene element (e.g. calling dirt "paved") the critic dutifully enforces that error, and the loop compounds the mistake in later clips. While this usually causes only gradual content drift, integrating a simple verifier to check object counts or action labels before rewriting could stop propagation of obvious prompt errors.

E. Qualitative Examples

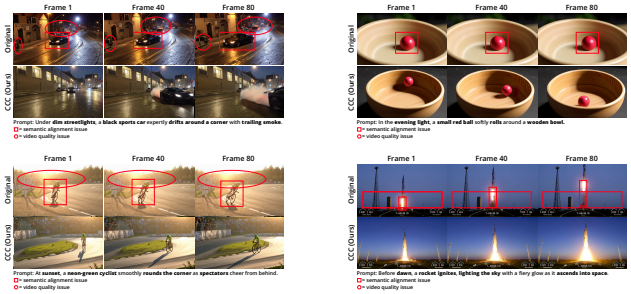

Figure 7: (a)–(d) original vs. CCC-refined videos. Each subfigure shows frames 1/40/80 from the original (top row) and our method (bottom row), with semantic-alignment issues marked in red squares and video-quality glitches in red circles.

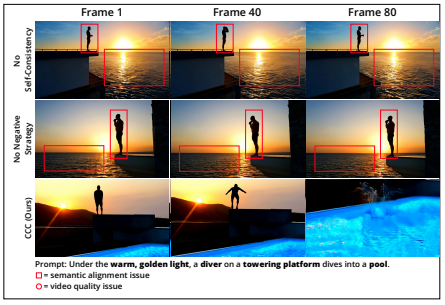

Figure 10: **Comparison with Ablations.** For the diver prompt, `Critique Coach Calibration` keeps the subject, motion, and scene intact, whereas ablations miss actions or key objects.

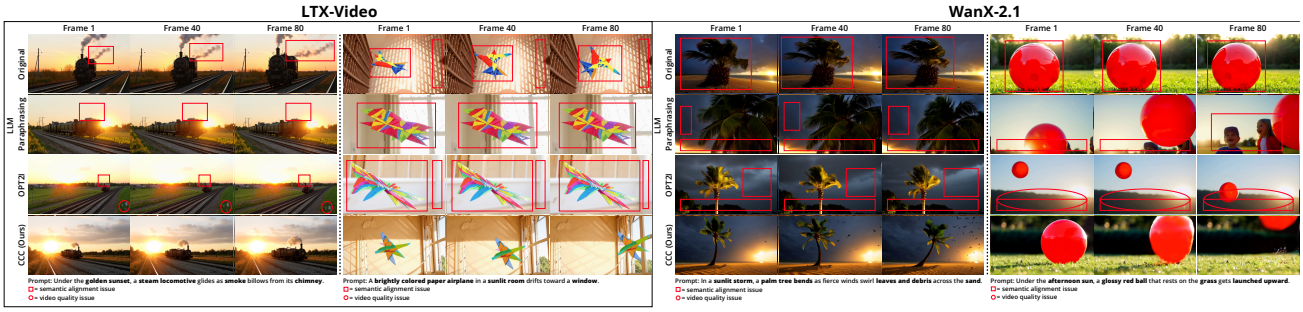

Figure 8: **Human preference scores (CCC vs. Origin) on Qwen-2.5-7b:** CCC consistently improves both semantic alignment and visual quality.

Figure 9: **Adding a fault-seeking clause ("There could be issues…") draws richer and more accurate critiques from the MLLM.** Across two generated videos, the negative-prompt strategy surfaces missing details that the neutral prompt overlooks, shifting responses closer to human ground truth.