CaC: Advancing Video Reward Models via Hierarchical Spatiotemporal Concentrating

Jiyuan Wang[1,2,*] **Huan Ouyang**[3,4,*] **Jiuzhou Lin**[4] **Chunyu Lin**[1,†] **Dewen Fan**[4,‡]
Boheng Zhang[4] **Haonan Fan**[4] **Fei Zuo**[4] **Jia Sun**[4] **Huaiqing Wang**[4]
Honglie Wang[4] **Yiyang Fan**[4] **Zhenlong Yuan**[4] **Zijun Li**[4] **Yongrui Heng**[4]
Guosheng Lin[2] **Fan Yang**[4,†] **Tingting Gao**[4]

[1]BJTU [2]NTU [3]BUPT [4]Kuaishou Technology

[*]Equal contribution [†]Corresponding author [‡]Project leader

Abstract

Recently, video generation models have achieved remarkable progress in global visual fidelity. However, when synthesizing complex dynamic content, they frequently produce **sparse subtle anomalies**. Existing video reward models primarily emphasize overall visual quality and text alignment; yet as generation quality improves, these coarse-grained defects become less frequent, while the sparse structural, temporal, and physical anomalies become an increasingly important bottleneck. Therefore, in this paper, we propose Concentrate and Concentrate (CaC), a coarse-to-fine anomaly reward model based on Vision-Language Models. During inference, it first conducts a global temporal scan to anchor anomalous time windows, then performs fine-grained spatial grounding within the localized interval, and finally derives robust judgments via structured spatiotemporal Chain-of-Thought reasoning. To equip the model with these capabilities, we construct the first large-scale generated video anomaly dataset with per-frame bounding-box annotations, temporal anomaly windows, and fine-grained attribution labels. Building on this dataset, we design a three-stage progressive training paradigm. The model initially learns spatial and temporal anchoring through single- and multi-frame supervised fine-tuning, and then is optimized by a reinforcement learning strategy based on two-turn Group Relative Policy Optimization (GRPO). Beyond conventional accuracy rewards, we introduce Temporal and Spatial IoU rewards to supervise the intermediate localization process, effectively guiding the model toward more grounded and interpretable spatiotemporal reasoning. Extensive experiments demonstrate that CaC can stably "concentrate" on subtle anomalies, achieving a 25.7% accuracy improvement on fine-grained anomaly benchmarks and, when used as a reward signal, CaC reduces generated-video anomalies by 11.7% while improving overall video quality.

1 Introduction

Recent advances in video generation models [5, 43, 45, 25, 17, 32] have significantly improved the visual fidelity of generated videos. However, when producing complex and dynamic content, these models still frequently exhibit hallucinations [20, 21], including structural distortions [16, 59], physical-law violations [61, 6], temporal inconsistencies [4, 36], and corrupted text rendering [31, 12]. Unlike the well-defined anomalies observed in other areas, anomalies in generated videos are far more complex and often manifest as **subtle** violations of physical/logical priors.

Recently, the reinforcement learning (RL) paradigm has significantly enhanced the capabilities of generative models by employing reward models to evaluate outputs and drive optimization. In the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: Anomalies in generated videos are highly sparse in both space and time — deformed shoes/legs may appear in only a few frames within a small region. Our CaC model temporally "Concentrates" on the anomalous frames (red box), and then spatially "Concentrates" on the distorted region (orange box), enabling precise and interpretable anomaly detection.

domain of video generation, there is a critical need for a robust reward model capable of providing stable and accurate feedback on video anomalies, thereby advancing video generation quality.

While existing video reward models [58, 56, 60, 64, 19, 54] play a pivotal role in aligning human preferences, they still largely emphasize visual quality and text alignment, leaving structural distortions and localization errors relatively under-explored. Unlike global visual quality issues, hallucinations in generated videos are often spatiotemporally sparse, appearing only in a few frames or occupying small localized regions. Consequently, identifying these subtle errors requires fine-grained spatiotemporal perception to isolate weak anomaly signals from large amounts of normal context, yet these faint signals are often overwhelmed by current models' global video-level reasoning. To bridge this gap, video reward modeling requires a dedicated mechanism that can actively concentrate on these sparse anomalies through spatiotemporal focusing, filtering out the vast normal context to enable fine-grained reasoning.

In this paper, we propose **Concentrate and Concentrate (CaC)**, a VLM-based reward model with a coarse-to-fine two-turn localization strategy for focusing on subtle anomalies. During inference, CaC first conducts a global scan to anchor the anomalous temporal range, followed by fine-grained spatial grounding within the selected time window. This design not only reduces the burden of spatiotemporal search, but also mitigates VLM hallucinations by deriving final judgments from interpretable spatiotemporal CoT reasoning.

To enable this coarse-to-fine reasoning ability, we construct a three-stage training framework. First, we train the model to recognize and localize anomalies. Given single frames extracted from videos, the model learns anomaly comprehension and bbox grounding, laying a foundation for subsequent video-level detection. Second, we extend training to multi-frame inputs, enabling the model to capture temporal errors in dynamic scenes. This stage jointly trains global video scanning and fine-grained video-clip localization, ensuring that the model outputs the formats required for two-turn interaction. Finally, we employ two-turn GRPO to optimize the VLM's reasoning trajectory, enabling it to produce more stable and accurate judgments. We design a diversified reward mechanism, including format verification, temporal IoU, spatial IoU, attribution IoU, and result accuracy, which guides the model toward more accurate spatiotemporal localization and interpretable anomaly judgment.

To support the above pipeline, we construct the first large-scale dataset for sparse anomalies in generated videos, with expert annotations of temporal segments, anomaly types, and **per-frame bboxes**. Although labor-intensive, these annotations are crucial: 1) they improve CoT ground-truth (GT) construction by allowing video clips with GT spatiotemporal annotations to be fed into the foundation model, effectively reducing hallucinations during label generation compared with prior work [19, 58]; 2) they provide references for attribution and temporal/spatial IoU rewards in the third training stage, guiding the model toward more interpretable reasoning through explicit localization.

Extensive experiments on our anomaly benchmark demonstrate that CaC reliably detects sparse anomalies, achieving **81.7%** overall accuracy and improving over the strongest baseline by 25.7%.

On a harder split with increasingly sparse anomalies, CaC's advantage grows as anomalies become shorter in duration and smaller in spatial extent. CaC also improves accuracy by 19.9% on the external Skyra-derived subset, providing an out-of-domain generalization. More importantly, when used as a reward signal for downstream video generation, CaC reduces generated-video anomalies by 11.7% and improves overall video quality, demonstrating that its judgments provide actionable optimization feedback. Our contributions can be summarized as follows:

- We construct the first large-scale comprehensive dataset for sparse anomalies in generated videos, providing anomaly types, reasons, temporal windows, and per-frame bboxes. These annotations support high-quality CoT synthesis for SFT and provide verifiable reward signals for the RL stage.
- We propose CaC, a coarse-to-fine two-turn reward model that recognizes sparse video anomalies through hierarchical spatiotemporal focusing. Its structured judgments can be converted into reward feedback for reducing anomalies in generated videos.
- We develop a three-stage progressive pipeline for CaC training. In the RL stage, we further introduce temporal, spatial, and attribution IoU rewards to guide the VLM toward grounded, interpretable anomaly CoT reasoning and stable judgments.

2 Related Work

Video Reward Models. With the development of generative models [50, 49, 51, 52], reward modeling has become a key technique for aligning them with human preferences. Early single-scalar approaches [63, 24, 33] are insufficient for capturing multi-faceted visual quality. Subsequent video-domain methods incorporate human-annotated ratings [26, 18] and Bradley–Terry loss [30, 35], with some frameworks unifying cross-task evaluation [57] or extending to multi-dimensional scoring [19, 55, 30]. More recent VLM-based reward models further integrate CoT reasoning [56], dynamic sampling, or visual rechecking [58, 54] to improve evaluation depth. However, localized structural and physical anomalies can still be overwhelmed by predominantly normal context. In contrast, CaC introduces a coarse-to-fine spatiotemporal grounding mechanism, effectively concentrating on sparse anomalies amid normal context.

Reinforcement Learning. The integration of RL [28] into LLMs and MLLMs [10, 53, 2] has significantly advanced their reasoning capabilities [62, 34, 42]. While early implementations relied on PPO [37], GRPO [38, 15] simplifies advantage estimation via group-relative baselines and leverages verifiable rewards to elicit long-chain reasoning without extensive preference data. Recently, GRPO has been further applied to visual understanding [57, 64, 54, 56, 40] and visual generation [65, 47, 29]. In this work, we use LLM-style GRPO to train the CaC reward model, while downstream video alignment follows generation-style GRPO.

Video Anomaly Detection. This traditional CV task primarily studies abnormal events in real-world surveillance videos, such as human, vehicle, and environment anomalies [1]. This setting is fundamentally different from our generated-video anomalies studied here: our anomalies refer to structural distortions, temporal inconsistencies, etc. A few recent benchmarks and evaluators study these anomalies in generated videos [21, 3, 58], but they typically focus on video-level judgments or rule-level assessment rather than fine-grained spatiotemporal anomalies. Additionally, detector-style methods [40, 27] discriminate real from AI-generated videos via artifact reasoning. Though their default task differs from ours, we still include them as baselines and leverage their data for out-of-domain evaluation given the scarcity of open-source generated-video artifact models.

Visual Focusing. This line of work aims to allocate computation to informative regions or moments for fine-grained visual understanding [48, 14]. Existing studies explore image-level zooming and cropping [7, 39], frame-level keyframe selection for long-video understanding [70, 71], and spatiotemporal grounding as a reference basis [46, 44]. These works improve general visual understanding or target grounding, but typically stop at perception or localization. In contrast, CaC focuses on sparse anomalies in generated videos and, after temporal focusing and spatial grounding, further converts the anomaly evidence into actionable reward feedback for improving video generation quality.

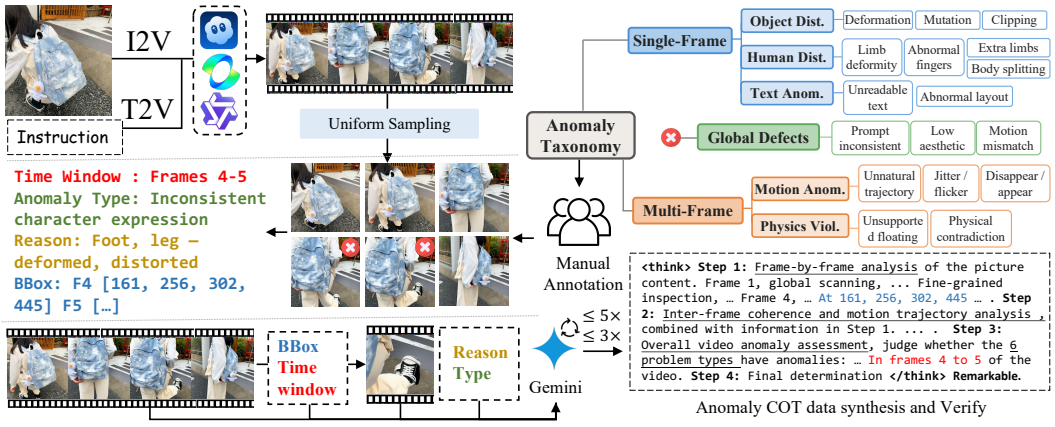

Figure 2: The processing pipeline for our CaC dataset. Please refer to Sec. 3.1 for details.

3 Method

3.1 Dataset Preparation

To enable the model to detect subtle hallucinations in generated videos, we construct a large-scale dataset with fine-grained spatiotemporal anomaly annotations, specifically targeting sparse video distortions. The overall construction pipeline is illustrated in Fig. 2.

Data Collection and Taxonomy. We collect images and prompts describing complex scenes and motions, and synthesize videos using SoTA video generation models, including Kling [43], Sora [5], and Wan [45]. We manually screen the outputs and curate ∼25K videos covering broad anomaly types. As shown in the top-right of Fig. 2, we build a new anomaly taxonomy designed to help VLMs identify subtle anomalies and support the subsequent three-stage training. Single-frame recognizable anomalies (Stage 1) include object distortion and human distortion, with common subtypes such as object deformation and extra limbs. We also include character anomalies, which are largely overlooked by existing reward models. Multi-frame anomalies (Stage 2) include motion anomalies and physical violations, covering common issues such as flickering and object disappearance. Moreover, we further filter these anomalous videos by excluding global quality issues such as text-video mismatch and low aesthetic quality, retaining 21,396 videos, ∼80% of the screened videos. This highlights that localized and sparse spatiotemporal anomalies are becoming increasingly dominant failure modes in current video generation.

CoT Synthesis with Human Annotation. High-quality CoT data can improve the effectiveness of reward models [19]. To construct such data for sparse video anomalies, we design an annotation pipeline that combines expert human labels with VLM-based CoT synthesis, extending prior practices [40, 41, 56]. As shown in Fig. 2 (left), we uniformly downsample each video into frame sequences, which are examined frame by frame by human experts. For anomalous samples, annotators label the temporal segments, anomaly types, underlying reasons, and affected-region bboxes. Given that subtle anomalies are difficult to annotate accurately, we organize 50 video evaluation experts, including 40 annotators and 10 reviewers, and adopt a two-round review mechanism where errors are returned for correction. We further randomly check 10% of the annotations to verify bbox and anomaly-type accuracy. Detailed annotation procedures are provided in Appendix B.1. Based on these annotations, as shown in Fig. 2 (bottom), we feed the original video, human-annotated GT, and cropped anomalous frames into Gemini-2.5-Pro [11], and design a four-step reasoning template to encourage comprehensive VLM analysis. Since some sparse anomalies are extremely difficult for VLMs to recognize, we introduce a generate-verify process, where the generated CoT and human annotations are sent back to Gemini to check their consistency. Samples verified within three rounds are deemed **salient anomalies**; those verified within five rounds are deemed **non-salient anomalies**; and the remaining samples are discarded, as they may be overly difficult or affected by human annotation errors. Ultimately, we retain 15,374 anomalous training samples. The detailed prompt templates are provided in Appendix B.2.

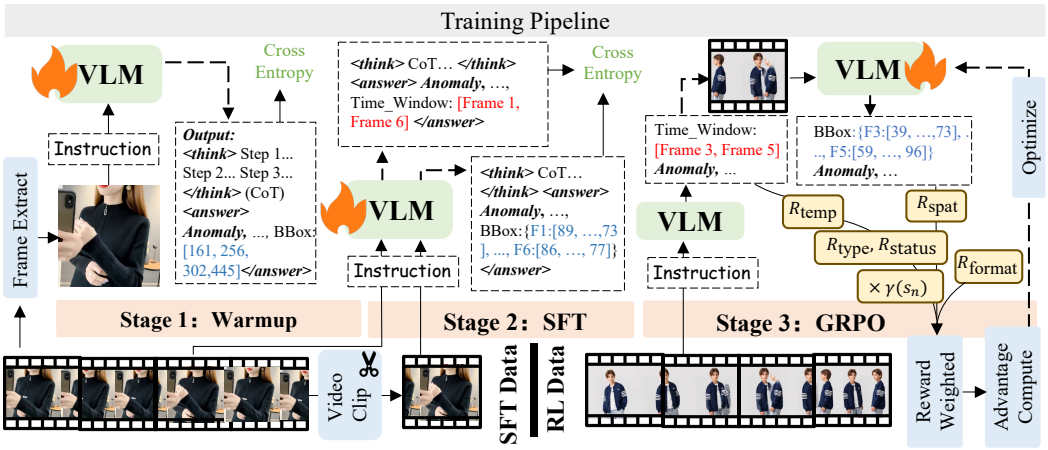

Figure 3: Overview of the CaC Reward Model's Coarse-to-Fine three-stage training paradigm. Please refer to Sec. 3.2 for details.

We further pair the anomalous set with an equal number of normal videos constructed using the same pipeline, resulting in a dataset of approximately 30K videos. More dataset details are provided in Appendix B.3, with additional statistics visualized in Fig. S1.

3.2 Multi-stage Progressive Training for Hierarchical Concentrating

Based on the constructed dataset, we build a 3-stage progressive training pipeline, as shown in Fig. 3.

Stage 1: Single-frame Warmup. Existing pre-trained VLMs are typically unexposed to anomaly scenes in generated videos. We therefore first teach the VLM to recognize and localize these distortions from anomalous frames. As shown in Fig. 3 (Stage 1), each training sample contains an extracted anomalous frame, the video prompt, and the corresponding CoT reasoning sequence constructed in Sec. 3.1. This sequence includes the judgment, reasoning, and spatial bbox coordinates of the affected region. Through this stage, the model acquires a structured CoT format and establishes a mapping from visual cues to anomaly attribution and spatial grounding.

Stage 2: Multi-frame SFT. After establishing frame-level anomaly understanding and spatial grounding capabilities, we move to video-level training. As shown in Fig. 3 (Stage 2), we propose a joint training strategy that prepares the model for both global temporal scanning and fine-grained spatial localization in the subsequent GRPO stage. Specifically, the model is optimized simultaneously on two types of video inputs: uniformly sampled (4 fps) global videos for localizing anomalous temporal regions within the full context, and densely sampled (8 fps) clips for precise anomaly localization after temporal focusing. These clips are cropped from the full videos as pure abnormal/normal temporal windows, and are all shorter than 1s. Using the constructed CoT supervision (Sec. 3.1), the model learns the spatial and temporal CoT reasoning format required by the third-stage two-turn video input and acquires basic video anomaly recognition capability.

Stage 3: Two-Turn GRPO. To further strengthen the model's multimodal spatiotemporal reasoning and improve the reliability of its CoT process, we adopt two-turn GRPO fine-tuning. As shown in Fig. 3 (Stage 3), given an instruction c and a video f, the model generates a structured anomaly prediction y in a predefined format. The prediction contains two rounds: in the first round, the model takes the uniformly sampled video f and predicts the anomaly judgment, temporal boundaries, and anomaly type, denoted as y^1. For each rollout, the video is then cropped along its predicted time window to obtain a localized clip $f_i' = \text{crop}(f, y_i^1)$. In the second round, the model takes the densely sampled f' together with the CoT derived from y^1, and predicts the anomaly judgment, spatial bboxes, and anomaly type, denoted as y^2. The complete response is defined as $y_i = [y_i^1, y_i^2]$.

During training, for each input (c, f), GRPO samples a group of G rollouts $\{y_1, y_2, \ldots, y_G\}$ from the old policy (VLM). Each y_i is evaluated by a weighted combination of rewards, and the rewards are normalized within the group to obtain the advantage $\hat{A}_i$. The policy is optimized by maximizing:

$$J_{\text{GRPO}}(\theta) = J_{\text{clip}}(\theta) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{1}$$

where π_θ and π_{ref} denote the fine-tuned and original VLMs, respectively, $J_{\text{clip}}(\theta)$ is computed from $\hat{A}_i$, and β controls the KL strength. See Appendix C.1 for details and the formal definition of GRPO. We next introduce these rewards.

3.3 Reward for CaC

Format Reward evaluates the structural correctness of outputs. We require each response to be a parsable JSON object satisfying the stage-specific, status-conditioned schema: valid judgment for all samples; empty anomaly list if predicted normal; and anomaly taxonomy with a valid time window or frame-indexed bboxes in the first/second turn if predicted anomalous. The format reward is defined as:

$$R_{\text{fmt}}(y_i) = \frac{1}{2}\sum_{t=1}^{2}\begin{cases}0, & \text{Valid}_t(y_i^t),\\ -1, & \text{otherwise.}\end{cases} \tag{2}$$

Invalid or unparsable outputs are penalized, while semantically incorrect predictions cannot receive additional positive reward merely by being well formatted.

Status Reward evaluates anomaly judgment accuracy by comparing the predicted status with the GT status. Similar to $R_{\text{fmt}}(\cdot)$, it assigns 0 to correct status predictions and -1 otherwise. For normal videos, a correct prediction can directly pass the status check in the first turn, while a false prediction is penalized and cannot receive IoU-based rewards because no GT anomaly exists. For anomalous videos, the model passes the status check only when its predictions are correct across both turns. Although anomalous samples can receive additional IoU-based rewards, these rewards are gated by correct multi-turn prediction and accurate attribution/localization; simply predicting "abnormal" therefore brings no benefit. For GT-normal samples, IoU terms are masked out, and GRPO normalizes rollouts per input, avoiding cross-sample bias.

Attribution Type IoU Reward evaluates the model's ability to identify five anomaly types introduced in Sec. 3.1 (object/human distortions; text/motion anomalies; physics violations), guiding the model to reach its judgment through accurate attribution CoT. For anomalous samples, we extract the anomaly type sets from the t-th turn prediction and the GT annotations, denoted as $\mathcal{T}^t_{\text{pred}}$ and $\mathcal{T}_{\text{gt}}$, and compute their set IoU:

$$R_{\text{type}}(y_i) = \frac{1}{2}\sum_{t=1}^{2}|\mathcal{T}^t_{\text{pred}} \cap \mathcal{T}_{\text{gt}}|/|\mathcal{T}^t_{\text{pred}} \cup \mathcal{T}_{\text{gt}}| \tag{3}$$

Temporal IoU Reward evaluates the accuracy of the temporal intervals identified in the first turn. Let $\mathcal{F}_{\text{pred}}$ and $\mathcal{F}_{\text{gt}}$ denote the sets of predicted and GT anomalous frame indices, respectively. The temporal reward is computed as:

$$R_{\text{temp}}(y_i) = |\mathcal{F}_{\text{pred}} \cap \mathcal{F}_{\text{gt}}|/|\mathcal{F}_{\text{pred}} \cup \mathcal{F}_{\text{gt}}|. \tag{4}$$

Spatial IoU Reward evaluates the precision of the frame-level bboxes predicted in the second turn. We compute spatial IoU on temporally matched frames, *i.e.,* $\mathcal{F}_{\text{pred}} \cap \mathcal{F}_{\text{gt}}$, where M is the number of GT bboxes on these frames. The reward is formulated as the mean IoU:

$$R_{\text{spa}}(y_i) = \frac{1}{M}\sum_{m=1}^{M}\text{IoU}(b_m^{\text{pred}}, b_m^{\text{gt}}),\ M>0;\quad 0,\ \text{otherwise.} \tag{5}$$

Together, these temporal and spatial IoU rewards encourage more accurate temporal grounding and spatial localization, leading to more reliable and interpretable CoT reasoning.

Reward Aggregation. In Sec. 3.1, each GT anomaly receives a saliency label $s_n \in \{\text{salient}, \text{non-salient}\}$ based on its verification difficulty. For anomalies with varying recognition difficulty, it is unreasonable to penalize all prediction errors equally. Therefore, we introduce a saliency discount factor $\gamma(s_n)$ and define the final reward as:

$$R(y_i) = w_1 R_{\text{fmt}}(y_i) + w_2\bar{\gamma}_i R_{\text{stat}}(y_i) + m_i\big(w_3\widetilde{R}_{\text{type}}(y_i) + w_4\widetilde{R}_{\text{temp}}(y_i) + w_5\widetilde{R}_{\text{spa}}(y_i)\big), \tag{6}$$

where m_i masks IoU terms for GT-normal samples, set to 1/0 for GT-anomalous/normal samples, respectively. The w_i values denote the weights of the respective reward components, and we set $(w_1, w_2, w_3, w_4, w_5) = (1, 2, 2, 2, 5)$. $\bar{\gamma}_i$ is set to 1 for GT-normal samples and otherwise denotes the average saliency discount over GT anomalies in the sample, and $\widetilde{R}_k = 1 - \bar{\gamma}_i(1 - R_k)$ for $k \in \{\text{type}, \text{temp}, \text{spa}\}$. We set $\gamma(s_n) = 0.5/1$ for non-salient/salient anomalies, respectively.

3.4 Two-turn Inference Strategy for Hierarchical Concentrating

Existing video reward models typically sample frames at fixed temporal intervals, feeding all sampled frames into the model simultaneously for inference. However, capturing subtle or fleeting anomalies inherently demands denser sampling, which incurs prohibitive computational overhead when processing long videos. In this work, we propose the Coarse-to-Fine Two-turn Inference strategy to efficiently localize anomalies.

First, the input video is uniformly downsampled to obtain a sparse frame sequence (4 fps), which is fed into the VLM for global temporal scanning. The model initially detects anomalies in the video; if detected, it outputs the anomalous time window, along with its attribution analysis. Second, we extract the video clip corresponding to the predicted time window and densely resample it (8 fps). This clip is then fed back into the VLM for fine-grained spatial reasoning, where the model confirms the anomalies and outputs the anomaly attribution and spatial bboxes. The first turn therefore proposes a candidate anomalous window, while the second turn makes the final anomaly decision. If the second turn does not confirm the anomaly, the video is conservatively classified as normal.

Reward Interface. CaC is used as a reward model through a structured-to-scalar interface. Given a video v and its instruction p, CaC produces a structured response, and we convert the final judgment into a smooth scalar reward using the VLM's vocabulary probabilities. Specifically, we read the probabilities of "normal" and "abnormal", denoted as P_{norm} and P_{abn}, and define:

$$R_{\text{CaC}}(v, p) = P_{\text{norm}}/(P_{\text{norm}} + P_{\text{abn}}). \tag{7}$$

Thus, CaC assigns higher rewards to likely normal videos and lower rewards to likely anomalous ones. The structured CoT, type, time window, and bbox can serve as interpretable evidence.

4 Experiment

Implementation. We adopt Qwen3-VL-8B [2] as the foundational model for CaC. All training stages unfreeze the full pipeline, including the ViT encoder, the adapter, and the LLM. In the warmup stage, we fine-tune the model with LoRA (rank=32) at a learning rate of 1×10^{-4}. In the Multi-frame SFT stage, we switch to full-parameter tuning with the same learning rate. In the RL stage, we perform two-turn GRPO for 250 steps with a learning rate of 1.0×10^{-6}, a rollout group size of $G = 8$, a batch size of 64, and gradient accumulation over 32 steps. Further implementation details are provided in Appendix C.2.

Benchmark. Existing benchmarks for video reward models [23, 30] primarily evaluate aesthetics and text-video alignment via pairwise preference accuracy, differing from our task in both evaluation dimensions and protocols. We therefore construct **CaC-Bench**. This evaluation set is separated from the training data in videos, prompts, anomaly distribution, and fine-grained taxonomy, and contains two subsets: (1) **CaC-Bench-Main** (500 normal + 500 anomalous videos), human-annotated with anomaly types and causes, all cross-verified; (2) **CaC-Bench-Hard** (400 anomalous-only videos), categorized by temporal duration ($<$1s / $\geq$1s) and spatial extent ($<$20% / $\geq$20% of frame area) into four difficulty levels to stress-test model sensitivity. We also evaluate on an external annotated subset derived from open-source Skyra/ViF-CoT-4K data [27]. More benchmark construction details are provided in Appendix A.

Baseline. On CaC-Bench-Main, we compare CaC with three groups of baselines: (1) leading general VLMs (Gemini-3.1-Pro [13], Gemini-2.5-Pro [11], GPT-4o [22], Qwen3-VL-8/32B [2]); (2) video reward models (UnifiedReward [57], UnifiedReward-Think [56], and VR-Thinker [54]); and (3) AIGC video detectors (VideoVeritas [40], Skyra-RL [27]). We keep frame sampling and prompting/protocol conversion consistent across methods, with details in Appendix C.2. For CaC-Bench-Hard, which contains only anomalous videos, we exclude models that exhibit strong prediction bias on CaC-Bench-Main (*e.g.,* Gemini-2.5-Pro variants that default to "abnormal") and retain only those with reasonably balanced detection behavior for a fair comparison.

4.1 Main Results

As shown in Tab. 1, CaC outperforms all existing methods in sparse anomaly detection, achieving an overall accuracy of 81.7% – a 25.7-point absolute improvement over the previous best. This not only demonstrates that CaC can accurately detect anomalies, but also confirms that prior reward models overlook this critical evaluation dimension. In contrast, Qwen3-VL-8/32B [2] achieve low recall

Table 1: Quantitative comparison against general MLLMs, video reward models, and video detectors on CaC-Bench-Main. Best and second-best results are in **bold** and underlined. Note that some baselines exhibit extreme single-sided metrics (*e.g.,* Skyra achieves 0.982 Anomalous Recall but only 0.018 Normal Recall), reflecting a strong prediction bias rather than detection capability.

Model	Anomalous Video			Normal Video			Acc.
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	
Gemini-2.5-Pro [11]	0.854	0.530	0.654	0.244	0.626	0.351	0.549
Gemini-3.1-Pro [13]	0.537	0.692	0.605	0.583	0.476	0.524	0.560
GPT-4o [22]	0.364	0.454	0.404	0.562	0.469	0.511	0.463
Qwen3-VL-8B [2]	0.108	0.574	0.182	0.920	0.508	0.654	0.514
Qwen3-VL-32B [2]	0.066	0.702	0.121	**0.972**	0.510	0.669	0.519
UnifiedReward [57]	0.538	0.563	0.550	0.582	0.557	0.569	0.560
UnifiedReward-Think [56]	0.428	0.504	0.463	0.578	0.503	0.538	0.503
VR-Thinker [54]	0.382	0.554	0.452	0.692	0.528	0.599	0.537
Skyra-RL [27]	**0.982**	0.500	0.663	0.018	0.500	0.035	0.500
VideoVeritas [40]	0.414	0.539	0.468	0.646	0.524	0.579	0.530
Ours (CaC)	0.910	**0.767**	**0.833**	0.724	**0.889**	**0.798**	**0.817**

Model	< 1s $< 20\%$	< 1s $\geq 20\%$	≥ 1s $< 20\%$	≥ 1s $\geq 20\%$
Gemini-3.1-Pro	0.21	0.46	0.31	0.57
UnifiedReward	0.32	0.53	0.37	0.51
VR-Thinker	0.23	0.39	0.40	0.33
CaC (Ours)	**0.70**	**0.69**	**0.76**	**0.71**

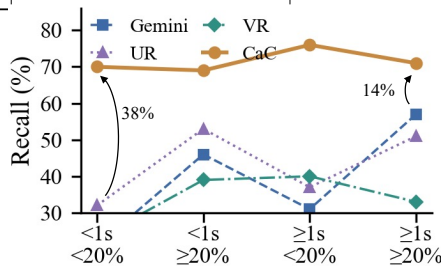

Figure 4: Quantitative comparison on CaC-Bench-Hard across four difficulty splits (we report recall because it contains only anomalous videos).

on anomalous videos, while Gemini-2.5-Pro [11] and Skyra-RL [27] exhibit low recall on normal videos, reflecting clear prediction biases: the former fail to recognize generative anomalies, whereas the latter are oversensitive.

As shown in Fig. 4, benefiting from the progressive coarse-to-fine localization design, CaC outperforms the best competing model on the four difficulty splits of CaC-Bench-Hard by 38%, 16%, 36%, and 14%, respectively. This advantage remains clear across different anomaly durations and spatial extents, and becomes pronounced for sparse anomalies, further validating the effectiveness of our approach. Per-category recall results are provided in Appendix D.1. CaC also maintains the strongest recall across all five anomaly categories (Tab. S3), indicating that the proposed reward model improves anomaly sensitivity without introducing diagnostic blind spots.

We further evaluate CaC on the open-source Skyra-ViF-CoT-4K data from an independent source (Appendix A.2). As shown in Tab. S1, CaC still improves over the strongest baseline by 19.9 points in accuracy. Qualitative results of the CaC CoT process are detailed in Appendix D.2.1 and Fig. S2, and Table S4 further reports Temporal IoU and Spatial IoU improvements over baselines. Together, these results demonstrate the effectiveness of the coarse-to-fine focusing strategy in detecting subtle video anomalies.

4.2 Video Alignment

To validate CaC as a reward model for video generation, we integrate it into two downstream optimization paradigms: Best-of-N sampling and Flow-GRPO on Wan-2.1-1.3B [45]. For Best-of-N, we generate five candidate videos for each prompt and select the one with the highest reward. For Flow-GRPO, CaC is used as a non-differentiable scalar reward for optimization, with group size 16 and LoRA rank 32; other details are provided in Appendix C.4.

Since CaC is a sparse-anomaly specialist, a general preference reward is still needed to evaluate text-video alignment and other holistic qualities, preventing reward hacking toward anomaly-free but low-quality videos. We therefore use UnifiedReward (UR) and average the normalized scores of UR and CaC as the final reward. We evaluate generated videos using VBench and expert anomaly annotations. As shown in Tab. 2, UR+CaC consistently improves over both the unaligned baseline and

Table 2: Quantitative comparison of reward models for improving video generation quality. We evaluate aligned videos with automatic VBench [21] metrics and human-expert anomaly annotations. Details of the selected metrics and qualitative before/after comparisons are provided in Appendices C.4 and D.2.2, respectively.

Group	Metric	Base model	Best-of-N		Flow-GRPO	
		Wan-2.1-1.3B	UR	UR+CaC	UR	UR+CaC
VBench	Dynamic Degree ↑	0.347	0.542	0.514	**0.597**	0.583
	Subject Consistency ↑	0.949	0.959	0.965	0.947	**0.970**
	Imaging Quality ↑	0.633	0.637	0.648	0.640	**0.667**
Expert	Anomaly Rate ↓	0.290	0.253	0.216	0.241	**0.173**

Table 3: Ablation results on CaC-Bench-Main. ⑥ *w/o Inference Strategy** is retrained with Stages 2 and 3 under a single-turn formulation. Due to space limitations, we report only Overall Accuracy; the remaining metrics are provided in Appendix D.3.

ID	Ablations	Overall Acc.	ID	Ablations	Overall Acc.
①	Base Model (Qwen3-VL-8B)	0.514	⑥	*w/o* Inference Strategy*	0.734
②	*w/o* Single-frame Warmup	0.701	⑦	*w/o* Temporal IoU Reward	0.629
③	*w/o* Multi-frame SFT	0.651	⑧	*w/o* Spatial IoU Reward	0.716
④	*w/o* Two-Turn GRPO	0.674	⑨	*w/o* Attribution IoU Reward	0.710
⑤	Base Model *w/* Inference Strategy	0.537	⑩	Ours (CaC)	**0.817**

UR-only alignment on subject consistency and imaging quality, while reducing the expert-annotated anomaly rate from 29.0% to 17.3% (an 11.7% absolute reduction). Dynamic degree shows only a marginal decrease, since static videos are inherently less prone to anomalies — this indicates that CaC does not trivially suppress motion to avoid anomalies, but instead provides complementary sparse-anomaly feedback without sacrificing overall generation quality.

Qualitative before/after comparisons are provided in Appendix D.2.2 and Fig. S3, showing that CaC mainly suppresses localized structural distortions and temporal inconsistencies present in the unaligned baseline.

4.3 Ablation Results

Effectiveness of the Progressive Training Paradigm. As shown in Tab. 3 ① *vs.* ②③④ *vs.* ⑩, all three training stages contribute substantially to the final performance and are jointly indispensable. In particular, ② removes the pre-training stage and reduces the overall accuracy by 12%, indicating that initial anomaly grounding provides an important foundation. ③ removes the SFT stage, which weakens the model's ability to output correct formats, makes the RL stage unstable, and reduces accuracy by 16%. ④ disables the RL stage and leads to a significant performance degradation with a 14% drop. Together, these results confirm the necessity of the progressive three-stage training paradigm.

Effectiveness and Efficiency of the Inference Strategy. As shown in Tab. 3 ⑤ *vs.* ①, when applying our two-turn CoT strategy to Qwen3-VL-8B, even in a training-free setting, it still marginally outperforms single-turn inference. In ⑥, CaC is retrained with a single-turn formulation that performs temporal and spatial localization simultaneously. Compared to ⑩, its accuracy drops by 8.3%. Taken together, both ⑤ and ⑥ validate the effectiveness of our inference strategy, suggesting that without the densified second-stage spatial grounding, the model struggles to reliably localize subtle anomalies.

Compute analysis is provided in Appendix C.3 and Table S2. We acknowledge that this two-turn strategy doubles the reward-model inference time, but this overhead remains acceptable in video RL training, as our downstream experiments show that video sampling is more than 8× slower than reward evaluation. Importantly, the added time cost yields a clear accuracy gain (for anomaly judgment, 56% accuracy is insufficient for reliable use), resulting in a substantially better performance-to-cost trade-off at the full training-system scale.

Effectiveness of Individual Rewards. As shown in Tab. 3 ⑦⑧⑨ *vs.* ⑩, we remove the Temporal/Spatial/Attribution IoU reward, respectively (the effectiveness of JSON-format/status rewards has been widely validated in prior work [57], so we omit them here). Among them, the Temporal IoU reward is the most critical: removing it causes a 19% performance drop because it directly

determines the decision range of the second turn. These ablations collectively show that using process-level rewards to guide the model toward high-quality spatiotemporal CoT reasoning leads to more accurate anomaly judgment.

5 Conclusion

In this paper, we presented CaC (Concentrate and Concentrate), a coarse-to-fine reward model for detecting and localizing subtle anomalies in generated videos. By combining a two-turn inference strategy, progressive three-stage training, and verifiable spatiotemporal rewards, CaC mitigates the attention dilution caused by the extreme sparsity of video anomalies and produces more reliable grounded judgments. We further introduced a dataset with per-frame bbox annotations to support both training and evaluation in this setting. Extensive experiments show that CaC consistently outperforms existing video reward models, detector-style baselines, and general VLMs, while downstream alignment results further confirm its value as an effective reward signal for improving video generation quality.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62573039 and U2441242, and by the "111 Center" (B25028).

References

[1] Ghazal Alinezhad Noghre, Armin Danesh Pazho, and Hamed Tabkhi. A survey on video anomaly detection via deep learning: Human, vehicle, and environment. *arXiv preprint arXiv:2508.14203*, 2025.

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

[3] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[6] Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, et al. Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. *arXiv preprint arXiv:2512.24551*, 2025.

[7] Miguel Carvalho, Helder Dias, and Bruno Martins. Cropvlm: Learning to zoom for fine-grained vision-language perception, 2026. URL `https://arxiv.org/abs/2511.19820`.

[8] Zeyang Chen, Mingnan Hu, and Bo Chen. Thermal image-guided complementary masking with multiscale fusion for multi-spectral image semantic segmentation. *Engineering Applications of Artificial Intelligence*, 150:110569, 2025.

[9] Zeyang Chen, Chunyu Lin, Yao Zhao, and Tammam Tillo. Unsupervised multi-modal domain adaptation for rgb-t semantic segmentation. *Computer Vision and Image Understanding*, page 104573, 2025.

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[12] Nikai Du, Zhennan Chen, Shan Gao, Zhizhou Chen, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025.

[13] Google. Gemini-3-pro. `https://deepmind.google/models/gemini/pro/`, 2025.

[14] Cheng Guan, Chunyu Lin, Zhijie Shen, Junsong Zhang, and Jiyuan Wang. Reper-360: Releasing perspective priors for 360° depth estimation via self-modulation. *arXiv preprint arXiv:2603.05999*, 2026.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[17] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

[18] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024.

[19] Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

[20] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025.

[21] 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*, pages 21807–21818, 2024.

[22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[23] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

[24] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

[25] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[26] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7793–7802, 2024.

[27] Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: Ai-generated video detection via grounded artifact reasoning. *arXiv preprint arXiv:2512.15693*, 2025.

[28] Youfang Lin, Jinji Fu, Haomin Wen, Jiyuan Wang, Zhenjie Wei, Yuting Qiang, Xiaowei Mao, Lixia Wu, Haoyuan Hu, Yuxuan Liang, et al. Drl4aoi: A drl framework for semantic-aware aoi segmentation in location-based services. *arXiv preprint arXiv:2412.05437*, 2024.

[29] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

[30] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[31] Zichen Liu, Yihao Meng, Hao Ouyang, Yue Yu, Bolin Zhao, Daniel Cohen-Or, and Huamin Qu. Dynamic typography: Bringing text to life via video diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14787–14797, 2025.

[32] Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

[33] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

[34] Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025.

[35] Pejaver V Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317): 194–204, 1967.

[36] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

[37] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[38] 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.

[39] Yuxiang Shen, Hailong Huang, Zhenkun Gao, Xueheng Li, Man Zhou, Chengjun Xie, Haoxuan Che, Xuanhua He, and Jie Zhang. Svfeye: A semantic-visual fusion framework with multi-scale visual context for multimodal reasoning, 2026. URL `https://arxiv.org/abs/2603.00171`.

[40] Hao Tan, Jun Lan, Senyuan Shi, Zichang Tan, Zijian Yu, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Videoveritas: Ai-generated video detection via perception pretext reinforcement learning. *arXiv preprint arXiv:2602.08828*, 2026.

[41] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *arXiv preprint arXiv:2503.20752*, 2025.

[42] Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[43] Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

[44] Xuezhen Tu, Jingyu Wu, Fangyu Kang, Qingpeng Nong, Kaijin Zhang, Chaoyue Niu, and Fan Wu. Bridging time and space: Decoupled spatio-temporal alignment for video grounding, 2026. URL `https://arxiv.org/abs/2604.08014`.

[45] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL `https://arxiv.org/abs/2503.20314`.

[46] Jiankang Wang, Zhihan Zhang, Zhihang Liu, Yang Li, Jiannan Ge, Hongtao Xie, and Yongdong Zhang. Spacevllm: Endowing multimodal large language model with spatio-temporal video grounding capability, 2025. URL `https://arxiv.org/abs/2503.13983`.

[47] Jin Wang, Jianxiang Lu, Guangzheng Xu, Comi Chen, Haoyu Yang, Linqing Wang, Peng Chen, Mingtao Chen, Zhichao Hu, Longhuang Wu, et al. Tagrpo: Boosting grpo on image-to-video generation with direct trajectory alignment. *arXiv preprint arXiv:2601.05729*, 2026.

[48] Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4976–4982. IEEE, 2024.

[49] Jiyuan Wang, Chunyu Lin, Lang Nie, Kang Liao, Shuwei Shao, and Yao Zhao. Digging into contrastive learning for robust depth estimation with diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4129–4137. ACM, 2024.

[50] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905*, 2025.

[51] JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025.

[52] Jiyuan Wang, Chunyu Lin, Lei Sun, Zhi Cao, Yuyang Yin, Lang Nie, Zhenlong Yuan, Xiangxiang Chu, Yunchao Wei, Kang Liao, et al. Geometry-guided reinforcement learning for multi-view consistent 3d scene editing. *arXiv preprint arXiv:2603.03143*, 2026.

[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[54] Qunzhong Wang, Jie Liu, Jiajun Liang, Yilei Jiang, Yuanxing Zhang, Jinyuan Chen, Yaozhi Zheng, Xintao Wang, Pengfei Wan, Xiangyu Yue, et al. Vr-thinker: Boosting video reward models through thinking-with-image reasoning. *arXiv preprint arXiv:2510.10518*, 2025.

[55] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024.

[56] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025.

[57] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

[58] Yuan Wang, Borui Liao, Huijuan Huang, Jinda Lu, Ouxiang Li, Kuien Liu, Meng Wang, and Xiang Wang. Thinking with frames: Generative video distortion evaluation via frame reward model. *arXiv preprint arXiv:2601.04033*, 2026.

[59] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7623–7633, 2023.

[60] Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025.

[61] Shang Wu, Chenwei Xu, Zhuofan Xia, Weijian Li, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Phyprompt: Rl-based prompt refinement for physically plausible text-to-video generation. *arXiv preprint arXiv:2603.03505*, 2026.

[62] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025.

[63] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[64] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

[65] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

[66] Zhenlong Yuan, Dapeng Zhang, Zehao Li, Chengxuan Qian, Jianing Chen, Yinda Chen, Kehua Chen, Tianlu Mao, Zhaoxin Li, Hao Jiang, and Zhaoqi Wang. DVP-MVS++: Synergize depth-normal-edge and harmonized visibility prior for multi-view stereo. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[67] Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. AutoDrive-R^2: Incentivizing reasoning and self-reflection capacity for VLA model in autonomous driving. In *The Fourteenth International Conference on Learning Representations*, 2026.

[68] Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, and Shuo Li. Video-STAR: Reinforcing open-vocabulary action recognition with tools. In *The Fourteenth International Conference on Learning Representations*, 2026.

[69] Zhenlong Yuan, Yue Wang, Dapeng Zhang, Kejin Cui, Rui Chen, Jing Tang, Lei Sun, Hongwei Yu, Chengxuan Qian, Xiangxiang Chu, Shuo Li, and Yuyin Zhou. What if agents could imagine? reinforcing open-vocabulary HOI comprehension through generation. *Advances in Neural Information Processing Systems*, 2026.

[70] Yuan Zhou, Litao Hua, Shilong Jin, Wentao Huang, and Haoran Duan. Reason: Reinforced causal search with information bottleneck for video understanding, 2025. URL https://arxiv.org/abs/2511.12530.

[71] Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding, 2025. URL https://arxiv.org/abs/2510.27280.

Appendix

In this appendix, we provide additional implementation details, experimental results, analyses, and discussions for a comprehensive evaluation and understanding of CaC. Detailed contents are listed as follows:

Contents

A Benchmark Details and Results **16**
- A.1 CaC-Bench Construction Details . . . 16
- A.2 External Evaluation on a Skyra Annotated Subset . . . 17

B Dataset Construction and Details **17**
- B.1 Detailed Annotation Guidelines . . . 17
  - B.1.1 Frame Extraction and Inspection . . . 17
  - B.1.2 Anomaly Structured Description . . . 18
  - B.1.3 Bounding Box Grounding . . . 18
- B.2 Prompt Templates for CoT Synthesis . . . 19
- B.3 Dataset Statistics and Visualizations . . . 19

C Additional Training and Inference Details **20**
- C.1 GRPO Formulation . . . 20
- C.2 Training Implementation Details . . . 21
- C.3 Inference Compute Analysis . . . 21
- C.4 Video Alignment Experiment Details . . . 21
  - C.4.1 VBench Metrics . . . 22

D Additional Experimental Results **22**
- D.1 Per-category Recall Results . . . 22
- D.2 Qualitative Results . . . 22
  - D.2.1 CaC Result Visualization . . . 22
  - D.2.2 Video Alignment Qualitative Results . . . 23
- D.3 Full Ablation Results . . . 23

E Limitations and Broader Impact **24**

Recent advances in learning-based visual perception and reasoning span multi-view stereo reconstruction, generative human–object interaction understanding, tool-assisted action recognition, and reasoning-based autonomous driving [66–69]. These studies provide broader context for research on visual perception and reasoning across diverse tasks.

A Benchmark Details and Results

A.1 CaC-Bench Construction Details

CaC-Bench is constructed as an independent evaluation suite for sparse anomaly detection in generated videos. It is separated from the training data in videos, prompts, anomaly distribution, and fine-grained taxonomy, and is never used for SFT, GRPO, prompt selection, reward-weight selection, or hyperparameter tuning. The goal is to evaluate whether a reward model can identify subtle localized structural, temporal, and physical anomalies under a test distribution that is distinct from the training stages. Specifically, unlike the data construction in Sec. 3.1, we collect a new set of 5K videos for CaC-Bench and do not require annotators to use the taxonomy in Fig. 2. Instead, GT labels are obtained through two rounds of consensus-based anomaly verification.

CaC-Bench-Main contains 1,000 generated videos, balanced between 500 normal and 500 anomalous samples. For anomalous samples, annotators record the anomaly type and the underlying cause, and the annotations are cross-verified before inclusion. Normal samples are also selected only when no clear localized structural, temporal, physical, or character anomaly is observed after two rounds of verification.

CaC-Bench-Hard contains 400 anomalous videos designed to stress-test sensitivity to sparse defects. We divide the samples into four difficulty groups according to two axes: temporal duration ($<$1s

Table S1: External evaluation on the Skyra-ViF-CoT-4K-derived annotated subset. Best and second-best results are in **bold** and underlined.

Model	Anomalous Video			Normal Video			Acc.
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	
Gemini-3.1-Pro [13]	0.164	0.689	0.265	**0.926**	0.526	0.671	0.545
UnifiedReward [57]	0.472	0.506	0.489	0.540	0.506	0.522	0.506
Qwen3-VL-8B [2]	0.448	0.526	0.484	0.596	0.519	0.555	0.522
Ours (CaC)	**0.564**	**0.881**	**0.688**	0.924	**0.679**	**0.783**	**0.744**

/ $\geq$1s) and spatial extent ($<$20% / $\geq$20% of the frame area). Since all videos in this split are anomalous, we report recall on each difficulty group. This split is used only for final evaluation and is not used to choose prompts, thresholds, reward weights, or model checkpoints.

A.2 External Evaluation on a Skyra Annotated Subset

To the best of our knowledge, there is currently no open-source benchmark specifically designed for generated-video anomaly detection. To provide an additional out-of-domain check beyond CaC-Bench and more fairly assess CaC's performance and generalization, we construct an auxiliary external evaluation subset from the open-source Skyra-ViF-CoT-4K data [27]. Skyra studies AI-generated video detection through grounded artifact reasoning, *i.e.*,, determining whether a video is AI-generated, and its annotated training data records human-perceivable visual artifacts in generated videos. These artifact categories are closely aligned with our anomaly taxonomy: person inconsistency corresponds to Human Distortion, text artifacts correspond to Character Anomalies, object inconsistency corresponds to Object Distortion, physics violations correspond to Physical Violations, and motion artifacts correspond to Motion Anomalies.

The Skyra-ViF-CoT-4K subset is not intended to replace CaC-Bench, since CaC-Bench focuses on sparse anomalies. Instead, it evaluates whether models can detect annotated artifacts from an independent data source while maintaining low false positives on normal videos. The subset is balanced between anomalous and normal samples, and all methods are evaluated under the same frame-sampling setting and binary normal/abnormal parsing protocol. It is worth noting that we do not use ViF-Bench because it only provides real/synthetic video labels and does not include anomaly annotations. We also do not compare against Skyra-RL because this auxiliary test subset is derived from its training set.

As shown in Table S1, existing general VLMs and other reward models remain close to random performance on this external subset, with either conservative normal-biased predictions or weak abnormal recall. In contrast, CaC achieves the best accuracy, anomalous F1, and normal F1, indicating that its anomaly-sensitive reasoning generalizes beyond CaC-Bench to artifacts annotated by an independent data source.

B Dataset Construction and Details

B.1 Detailed Annotation Guidelines

We provide the complete annotation protocol for the CaC dataset, which comprises four stages: **frame extraction and inspection**, **anomaly structured descriptions**, **bounding box grounding**, and **quality assurance**. We assemble a dedicated team of 50 video evaluation experts (40 annotators and 10 senior reviewers), all of whom complete a calibration session with exemplar videos covering our five predefined anomaly categories prior to formal annotation. The protocol operates on generated videos and does not rely on personal user metadata.

B.1.1 Frame Extraction and Inspection

As described in Sec. 3.1, each video is first uniformly downsampled to extract a sequence of frames. Human experts then inspect each frame sequentially to detect anomalies. This frame-level examination ensures that both single-frame defects and multi-frame anomalies are captured. Videos in which no anomaly is identified in any frame are labeled as normal samples.

B.1.2 Anomaly Structured Description

For each of the five predefined categories in our taxonomy, annotators determine whether the anomaly type is present via binary classification. A single video may be associated with multiple categories simultaneously. The five categories are: **Object Distortion** (deformation, mutation, or clipping artifacts in inanimate objects), **Human Distortion** (limb deformity, abnormal fingers, extra limbs, or body splitting in human subjects), **Motion Anomalies** (unnatural trajectories, jitter/flicker, or sudden disappearance/appearance of entities across frames), **Physical Violations** (unsupported floating, interpenetration, or other behaviors contradicting real-world physics), and **Character Anomalies** (unreadable text, abnormal layout, or inconsistent textual content across frames).

Annotation Format. For each detected anomaly, annotators provide a structured description that encodes the frame range, the affected objects, the anomaly types, and the underlying reason. The temporal boundaries are recorded as start and end indices on the uniformly sampled frame sequence, rather than timestamps. This unified format enables frame-level grounding of each anomaly instance while the underlying reason field captures the semantic context behind the anomalies. When multiple anomaly instances belong to the same category, they are concatenated within a single group; anomalies belonging to different categories are recorded in separate groups following the category ordering defined in the taxonomy. This hierarchical scheme enables unambiguous parsing of annotations into per-category anomaly groups, each containing one or more frame-grounded instances.

Annotation Conventions. To resolve ambiguities during annotation, we establish several guiding principles. When an anomaly is detectable from a single frame (*e.g.,*, a deformed hand or unreadable text), it is assigned to a single-frame category (Object Distortion, Human Distortion, or Character Anomalies) even if it persists across multiple frames, whereas multi-frame categories (Motion Anomalies and Physical Violations) are reserved for anomalies whose identification inherently requires cross-frame comparison. To distinguish Object Distortion from Physical Violations, the former captures structural deformation of the object itself, while the latter targets implausible interactions between objects or with the environment. For Character Anomalies, annotators compare textual content across the full frame sequence when needed, and inconsistent text is annotated with the complete frame range over which the inconsistency is observable, even if each individual frame appears plausible in isolation.

B.1.3 Bounding Box Grounding

For each identified anomaly instance, annotators draw a tight bounding box around the anomalous region in the corresponding frames within the annotated frame range. Bounding boxes are stored as $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ in the Qwen-style normalized coordinate system, where each frame axis is mapped to the range $[0, 1000]$. The following principles guide this process:

- Each bounding box must fully encompass the relevant distorted region such that the anomaly can be identified solely from the visual content within the box, without relying on surrounding context.
- Multiple bounding boxes may be assigned to a single anomaly instance when the corresponding distortion manifests in multiple disjoint regions within the same frame.
- When partial occlusion occurs, annotators approximate the full extent of the affected area based on visible cues and structural priors.
- Bounding boxes should be as tight as possible, avoiding unnecessary inclusion of irrelevant visual content to minimize interference from unrelated structures.
- For multi-frame anomalies that involve entity state transitions (*e.g.,*, sudden disappearance, appearance, or abrupt transformation), bboxes are annotated on the *boundary frames* where the transition occurs. Specifically, for disappearance events, the bounding box is placed on the last frame in which the entity is still visible, enclosing its full extent at that moment. For sudden appearance events, the bounding box is placed on the first frame in which the entity becomes visible. This convention anchors the bbox annotation to the most informative frame while the structured description captures the full frame range of the anomaly.

To guarantee annotation reliability, we adopt a rigorous **two-round** review mechanism. In the first round, each completed annotation is independently verified by a senior reviewer, who examines the anomaly types, frame boundaries, bboxes, and underlying reasons; any inconsistencies are flagged

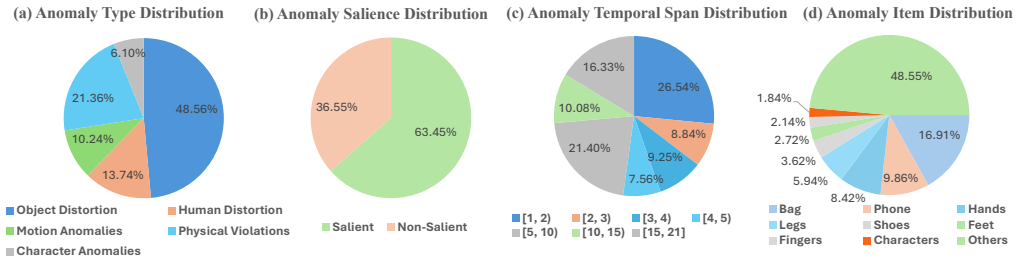

Figure S1: Statistical analysis of the CaC dataset.

and returned to the original annotator for correction, and this cycle repeats until the reviewer approves the annotation. In the second round, we randomly sample 10% of the accepted annotations for a spot-check audit conducted by a separate group of reviewers, verifying the accuracy and consistency of both the bboxes and the anomaly types. Only batches achieving an inter-annotator agreement rate above 95% are accepted into the final dataset; those falling below this threshold are returned for complete re-annotation.

B.2 Prompt Templates for CoT Synthesis

In this section, we provide a detailed overview of the prompt templates used throughout the CoT data construction process. As described in Sec. 3.1, we feed each video along with its manual annotations into Gemini-2.5-Pro [11] to synthesize structured reasoning chains. The core synthesis prompt is shown in Figs. S4 and S5. Specifically, we supply the annotated anomaly types, temporal spans, underlying reasons, and corresponding bboxes to guide the model toward the difficult regions while still requiring it to re-examine the visual evidence. The model is instructed to follow a four-step reasoning template: first performing detailed frame-level inspection, then analyzing cross-frame consistency and motion, then making an overall anomaly assessment with concrete attribution, and finally producing a structured JSON output with the anomaly status, frame range, label, reason, problem region, and per-frame bboxes. Although the ground-truth annotations are provided, they are treated as guidance rather than answers to be copied; the model must independently verify whether the visual content supports the annotated anomaly. This design reduces hallucinated rationales and yields higher-quality CoT labels for subtle anomalies that are difficult for VLMs to discover from the raw video alone.

For the verification stage, the generated CoT is checked against the human annotations to ensure that the reasoning chain is not only plausible but also aligned with the verified labels and localization; the verification prompt is shown in Fig. S6. For anomaly presence and type predictions, we perform exact string matching against the ground-truth labels to determine correctness; for the temporal segments and spatial bounding boxes, we compute the Intersection-over-Union (IoU) between the predicted and ground-truth regions, and consider a prediction correct if the IoU exceeds 0.5. A sample passes verification only when all criteria are simultaneously satisfied. This verification is necessary because GT-conditioned generation may still produce inconsistent reasoning, incorrect attribution, or imprecise localization, especially for non-salient anomalies. The process is repeated for up to five rounds—samples passing within the first three rounds are deemed salient anomalies, those passing in the fourth or fifth round are deemed non-salient anomalies, and those failing all five rounds are discarded to avoid introducing noise. By incorporating detailed descriptions of all five anomaly categories from our taxonomy into both prompts, we ensure that the synthesized CoT data faithfully reflects the fine-grained spatiotemporal reasoning required for video anomaly detection, thereby providing high-quality supervision for our three-stage progressive training pipeline.

B.3 Dataset Statistics and Visualizations

The CaC dataset contains 30,748 videos in total, evenly split between normal and anomalous samples. All videos are paired with synthesized CoT reasoning data. The anomalous videos are further enriched with manual annotations, including anomalous video segments, anomaly types, underlying reasons, corresponding bboxes, and generated saliency labels. **Quality Assurance.** As shown in Fig. S1, we report four complementary statistics of the anomalous subset: (a) anomaly type distri-

bution, (b) anomaly saliency distribution, (c) anomaly temporal span distribution, and (d) anomaly item distribution. In (a), Object Distortion accounts for the largest proportion, indicating that artifacts in inanimate objects remain the most common failure mode in current video generation models; Physical Violations are the second most frequent category, while Human Distortion, Motion Anomalies, and Character Anomalies appear with smaller proportions. In (b), most anomalous samples are classified as salient, while the remaining samples are non-salient. In (c), anomalous segments are usually short, confirming that many generation failures appear only within a few frames and can be overlooked by reward models that rely on holistic video-level assessment. In (d), the affected items are concentrated on fine-grained and small-scale entities, especially human body parts such as hands, fingers, legs, and feet, as well as objects such as bags and phones. Together, these statistics show that generated-video anomalies are sparse in both duration and affected region, motivating a coarse-to-fine reasoning strategy [8, 9] for reliable localization and judgment.

C Additional Training and Inference Details

C.1 GRPO Formulation

In this section, we provide the detailed formulation of the Group Relative Policy Optimization (GRPO) objective. For simplicity, we denote the overall conditioning context at step t as $X_{i,t}$, which encapsulates the task instruction c, the original or cropped video frames (f or f'), and the previously generated tokens $y_{i,<t}$ across the two autoregressive stages. In the second turn, the cropped frames are rollout-specific: each f_i' is constructed from the frame-index window predicted in the corresponding first-turn response y_i^1. If the predicted window is empty, invalid, or cannot be parsed into frame indices, we do not construct f_i' and the rollout terminates after the first turn.

For each input (c, f), the old policy samples a group of G complete two-turn rollouts $\{y_i\}_{i=1}^{G}$. Each rollout receives a scalar reward $R(y_i)$ from the weighted reward function in Sec. 3.3. We compute the group-relative advantage by normalizing rewards within the group:

$$\hat{A}_i = \frac{R(y_i) - \text{mean}(\{R(y_j)\}_{j=1}^{G})}{\text{std}(\{R(y_j)\}_{j=1}^{G}) + \epsilon_A}, \tag{8}$$

where ϵ_A is a small constant for numerical stability.

During training, the surrogate objective is maximized. The clipped surrogate objective $J_{\text{clip}}(\theta)$ restricts the policy update to prevent excessively large step sizes, and is formulated at the token level over all generated tokens in the rollout as:

$$J_{\text{clip}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(r_{i,t}(\theta)\hat{A}_i, \text{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right), \tag{9}$$

where $r_{i,t}(\theta)$ is the probability ratio between the current policy π_θ and the old policy $\pi_{\theta_{\text{old}}}$:

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid X_{i,t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid X_{i,t})}, \tag{10}$$

and ϵ is the clipping hyperparameter.

To prevent the fine-tuned model π_θ from deviating too far from the initial reference model π_{ref}, a KL divergence penalty is applied. Following standard GRPO implementations, we employ an unbiased estimator of the KL divergence. The KL penalty is then written as:

$$D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\left(\rho_{i,t} - \log\rho_{i,t} - 1\right), \rho_{i,t} = \frac{\pi_{\text{ref}}(y_{i,t} \mid X_{i,t})}{\pi_\theta(y_{i,t} \mid X_{i,t})}. \tag{11}$$

Combining these components yields the full GRPO objective:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{(c,f)\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[J_{\text{clip}}(\theta) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right], \tag{12}$$

where β is the coefficient controlling the strength of the KL penalty. By calculating advantages relative to the group rather than using an external value network, GRPO significantly reduces the memory overhead while maintaining stable optimization.

Table S2: Preliminary inference-time and accuracy comparison on a single-machine GPU setup. We report the average wall-clock time per video (21 frames sampled).

Method	Stage / Pass	Avg. Time ↓	Acc. ↑
Video Generation	Sampling (1 GRPO step)	272s	-
UnifiedReward	Single pass	**3.79s**	<u>0.560</u>
UnifiedReward-Think	Single pass with reasoning	<u>16.5s</u>	0.503
VR-Thinker	Multiturn adaptive stopping	42.3s	0.537
Gemini-3.1-Pro	API calling	65.6s	<u>0.560</u>
CaC (Ours)	Two-turn inference	33.2s	**0.817**

C.2 Training Implementation Details

We adopt Qwen3-VL-8B [2] as the backbone of CaC and train the full model pipeline across all stages, including the vision encoder, adapter, and language model. The three training stages use disjoint data splits, and each stage keeps a 1:1 ratio between normal and anomalous samples. The first stage learns single-frame anomaly grounding from about 20K frames. The anomalous frames are sampled from 5K videos that contain single-frame recognizable anomalies, while normal frames are randomly sampled from normal videos. The second stage performs multi-frame SFT with about 20K full-video samples and 15K local clips. Full-video samples are collected from the CaC dataset to train global temporal localization. Clip-level samples are cropped from the full videos and restricted to within 1 second. For anomalous videos, we require the cropped clip to fully cover the annotated anomalous region, because partial crops of cross-frame inconsistency events may introduce ambiguous or incorrect labels. The third stage constructs 7.5K training instances for two-turn GRPO from a non-overlapping data split.

For optimization and compute, the warmup stage uses LoRA with rank 32 and a learning rate of 1×10^{-4}, and is trained for 9,330 steps on 8 GPUs, taking about 11.5 hours. The multi-frame SFT stage switches to full-parameter tuning with the same learning rate, and is trained for 7K steps on 8 GPUs, taking about 30.5 hours. The RL stage runs GRPO with a learning rate of 1.0×10^{-6}, rollout group size $G = 8$, batch size 64, and gradient accumulation over 32 steps; it is trained for 250 steps on 16 GPUs, taking about 42 hours.

Baseline Evaluation Protocol. For all comparison methods, we first use the same prompt as CaC under the same frame-sampling setting to obtain a binary normal/abnormal decision. For reward and detector-style baselines, we additionally run their official inference prompts or scoring protocols described in the corresponding papers [57, 56, 40, 54, 27]. We report the stronger result between the unified anomaly-prompt evaluation and the official-protocol evaluation for each baseline.

C.3 Inference Compute Analysis

As shown in Table S2, compared with single-turn reward models, the two-stage strategy roughly doubles the reward-model inference time because the model is queried twice. This is a reasonable trade-off: CaC spends more inference time but improves the overall accuracy from the strongest baseline's 0.560 to 0.817, yielding a 25.7-point gain. This overhead is also modest in video-generation RL, where video sampling typically dominates reward evaluation. In a representative training iteration, video sampling takes at least 4 minutes, whereas reward-model evaluation takes at most 1 minute. Thus, the additional verification cost does not dominate the end-to-end optimization budget.

C.4 Video Alignment Experiment Details

For downstream video alignment, we use Wan-2.1-1.3B [45] as the base video generation model and evaluate two optimization paradigms. For Best-of-N sampling, we generate five candidate videos for each prompt and select the one with the highest reward score. For Flow-GRPO, CaC is used as a non-differentiable scalar reward and is combined with UnifiedReward by averaging the normalized reward scores, which preserves general alignment quality while discouraging localized structural, temporal, and physical artifacts. We use 240 prompts for training. Flow-GRPO is trained on 32 GPUs for 120 steps with batch size 2, rollout group size 16, LoRA rank 32, noise level 0.3, and 50 sampling steps. Each generated video is 480P, 5 seconds long, and contains 81 frames. During optimization, generated videos are scored by the same two-turn CaC inference framework as in

Table S3: Recall comparison of CaC with representative reward and detector-style baselines across different anomaly types. Best and second-best results are in **bold** and underlined.

Model	Object Dist.	Human Dist.	Motion Anom.	Physics Viol.	Character Anom.
Gemini-3.1-Pro	0.498	0.361	0.343	0.578	0.621
Qwen3-VL-32B	0.072	0.136	0.125	0.111	0.483
UnifiedReward-Think	0.162	0.157	0.106	0.223	0.474
VideoVeritas	0.245	0.176	0.171	0.198	0.572
Ours (CaC)	**0.704**	**0.492**	**0.414**	**0.633**	**0.724**

Sec. 3.4, using 4 fps in the first turn and 8 fps in the second turn. This CaC-based scoring is used only for reward selection or optimization; for the final evaluation in Tab. 2, VBench metrics are computed automatically, while the anomaly rate is obtained from human-expert annotations.

C.4.1 VBench Metrics

We report three VBench [21] metrics that are most relevant to generation-induced distortions and inconsistencies. Subject Consistency measures whether the main subject remains visually stable across frames, making it sensitive to identity drift, object deformation, and structural distortions. Imaging Quality evaluates frame-level visual degradation such as blur, noise, exposure errors, and low-level artifacts. Dynamic Degree serves as a necessary control metric: high consistency can be obtained by producing nearly static videos, so we use it to verify that the aligned model preserves nontrivial motion while improving subject stability and visual quality.

D Additional Experimental Results

D.1 Per-category Recall Results

To evaluate CaC's diagnostic sensitivity to different anomaly patterns, we compare its recall with representative baselines across the predefined anomaly categories. As shown in Table S3, CaC achieves the highest recall for all anomaly types, indicating that it can detect not only salient structural distortions but also motion anomalies and physical violations that require cross-frame reasoning. These results suggest that CaC provides stable anomaly coverage without obvious category-specific diagnostic blind spots.

D.2 Qualitative Results

D.2.1 CaC Result Visualization

Fig. S2 visualizes the complete two-turn reasoning process of CaC. In the first turn, the model performs a global temporal scan over sparsely sampled frames, examines suspicious anomalies, and outputs coarse temporal anchors. The flagged interval is then clipped and densely resampled (e.g., 4 fps $\rightarrow$ 8 fps), and in the second turn, the model conducts fine-grained spatial grounding within this narrowed window, producing structured diagnoses with anomaly labels, explanations, and bounding boxes. As shown in the examples, CaC successfully detects subtle physics violations such as a physically implausible floating phone and an object appearing from nowhere in the baby's hand — anomalies that are highly sparse and easily overlooked by holistic evaluation.

Table S4 further quantifies localization quality by reporting Temporal IoU and Spatial IoU for CaC and other baselines, showing that CaC substantially improves spatiotemporal anomaly localization accuracy.

Table S4: Temporal IoU and Spatial IoU comparison of CaC and baseline models.

Model	Temporal IoU	Spatial IoU
Gemini-3.1-Pro	0.163	0.085
Base Model *w/* Inference Strategy	0.235	0.067
Ours (CaC)	**0.695**	**0.340**



Figure S2: Qualitative analysis of CaC. **Bold text** marks key analysis, while underlined text indicates suspicious anomalies that are effectively filtered out. Redundant analyses of normal frames are omitted for clarity. These examples also illustrate that generated-video anomalies are highly sparse, appearing only in a few frames and occupying small local regions.

D.2.2 Video Alignment Qualitative Results

To qualitatively examine whether CaC provides useful feedback for video alignment, we compare generated videos before and after RL-guided optimization (using UR+CaC). As shown in Fig. S3, the baseline model exhibits various localized artifacts — train carriage distortion, an unnatural mouth opening in the dog, blurry faces, and temporally inconsistent road signs behind the motorcycle — all of which are substantially mitigated after optimization. These results confirm that incorporating CaC as an anomaly-sensitive reward signal effectively reduces localized structural defects and temporal inconsistencies while preserving visual coherence.

D.3 Full Ablation Results

Table S5 reports the complete ablation results with anomalous-video metrics, normal-video metrics, and overall accuracy.

Table S5: Complete ablation results on CaC-Bench-Main. ⑥ *w/o Inference Strategy** is retrained with Stages 2 and 3 under a single-turn formulation.

ID	Ablations	Anomalous Video			Normal Video			Acc.
		Recall	Precision	F1-Score	Recall	Precision	F1-Score	
①	Base Model (Qwen3-VL-8B)	0.108	0.574	0.182	<u>0.920</u>	0.508	0.654	0.514
②	*w/o* Single-frame Grounding Warmup	0.706	0.699	0.702	0.696	0.703	0.699	0.701
③	*w/o* Multi-frame SFT	0.506	0.713	0.592	0.796	0.617	0.695	0.651
④	*w/o* Two-Turn GRPO	0.600	0.704	0.648	0.748	0.652	0.696	0.674
⑤	Base Model *w/* Inference Strategy	0.102	**0.785**	0.181	**0.972**	0.520	0.677	0.537
⑥	*w/o* Inference Strategy*	<u>0.802</u>	0.706	<u>0.751</u>	0.666	<u>0.771</u>	0.715	<u>0.734</u>
⑦	*w/o* Temporal IoU Reward	0.484	0.682	0.566	0.774	0.600	0.676	0.629
⑧	*w/o* Spatial IoU Reward	0.610	<u>0.774</u>	0.682	0.822	0.678	<u>0.743</u>	0.716
⑨	*w/o* Attribution IoU Reward	0.668	0.729	0.697	0.752	0.694	0.722	0.710
⑩	Ours (CaC)	**0.910**	0.767	**0.833**	0.724	**0.889**	**0.798**	**0.817**

E Limitations and Broader Impact

CaC is designed as a specialist reward model for sparse structural, temporal, and physical anomalies in generated videos, and therefore should not be used as the only objective for downstream alignment. We have to combine CaC with UnifiedReward: the general reward preserves text-video alignment and holistic quality, while CaC provides complementary anomaly-sensitive feedback. However, we believe this is also a future direction: as video generation becomes increasingly realistic, reliable evaluation may require multiple expert reward models working together rather than a single monolithic reward.

Due to limited computational resources, our experiments are still restricted in scale. Although our training already uses 16-GPU and 32-GPU settings for the most expensive stages, broader multi-seed studies, larger downstream alignment sweeps, and evaluation on substantially longer videos would require considerably more compute. We therefore leave more exhaustive scaling analysis to future work.

The dataset construction also relies on fine-grained human annotations of frame ranges and bounding boxes. These annotations improve supervision quality, but they are costly and may still contain subjective ambiguity in borderline cases. We hope that releasing and further improving CaC can support future work on more automated fine-grained annotation pipelines, reducing the manual burden while preserving localization quality.

Finally, although the current RL formulation obtains a continuous reward through token-probability-based optimization, the reward feedback is still applied at the generated-video level. While CaC predicts temporal windows and spatial boxes, the optimization signal is not yet immediately mapped back to the exact anomalous frames or regions in the generator's denoising process. A promising future direction is to combine anomaly localization with more localized RL guidance, so that the predicted spatiotemporal boxes can directly indicate where and how generation should be corrected.

From a broader-impact perspective, CaC can support safer and more reliable video-generation systems by exposing subtle structural, temporal, and physical artifacts that are often missed by holistic evaluators. At the same time, improving anomaly detection and reward feedback may also help optimize generators toward more realistic synthetic media. Future release and deployment should therefore consider appropriate access control, usage restrictions, provenance tracking, and monitoring for harmful or misleading applications.

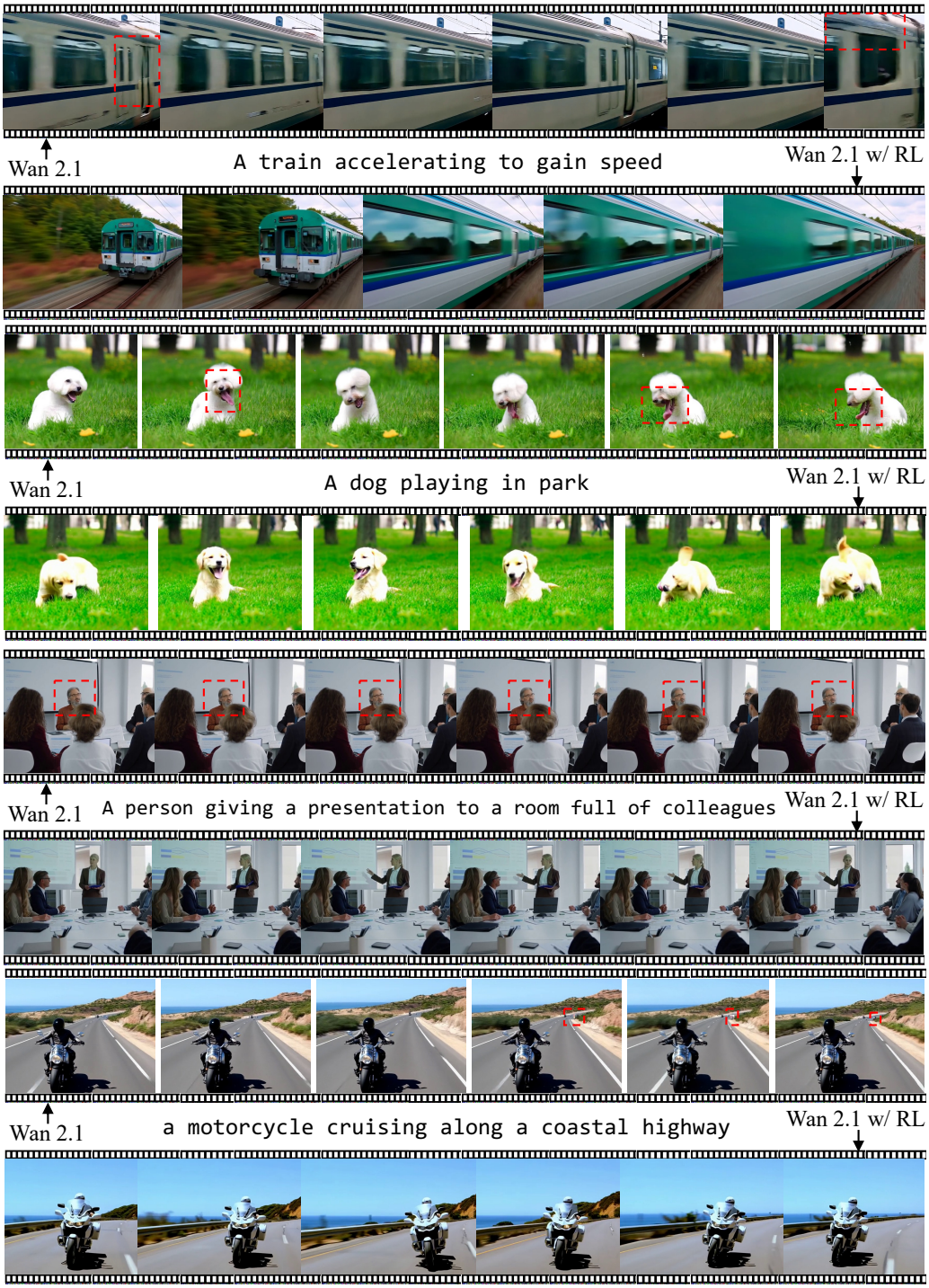

Figure S3: Qualitative comparison between videos generated by Wan-2.1-1.3B [45] before and after RL optimization with CaC. Red dashed boxes highlight anomalous regions in the baseline outputs.

Prompt for Video Anomaly Analysis

You are an expert in AI-generated video assessment, responsible for detecting whether anomalies exist in a video. You now need to analyze an AI-generated video.

Abnormal During the analysis process, please pay special attention to the following phenomena that may exist in the video: `{error_description}` .

Normal Human annotation considers this video to be normal, and you should conduct a careful review and inspection. Human annotation is generally correct. Of course, if you observe clear and explicit problems, you may also judge it as abnormal. But for minor anomalies, do not easily judge it as abnormal. However, if after careful observation you confirm that the human annotation is incorrect, you may adhere to your own judgment.

Abnormal The following information comes from human annotation and is only intended to help you localize the key observation region; it is not guaranteed to be correct. If after careful inspection you believe the whole snippet is normal and the human hint does not match the visual content, you may adhere to your actual observation and truthfully output "normal." This input is already a clipped anomalous time segment; if you finally judge it as abnormal, then the attributed time region is by default the entire snippet. During the analysis process, please pay special attention to the following phenomena that may exist in the video: `{error_description}` . The following are reference coordinates of the target region in each frame (`[xmin, ymin, xmax, ymax]`). When outputting BBOX, please use them as a reference and refine them according to the actual visual content: `{bbox_hint_str}` . The BBOX dictionary must cover every frame in these `{snippet_length}` frames and must not omit any frame.

Normal Human annotation considers this snippet to be normal, and you should conduct a careful review and inspection. If the video is overall smooth and natural, with no obvious visible defects, please judge it as normal. But if after careful observation you confirm that the human annotation is incorrect, and there is indeed a clear, explicit, and explainable anomaly in the video, you may still judge it as abnormal.

Abnormal However, in your thinking and response, you must behave as if you are completely unaware of the above hints. You must not mention in any form expressions such as "according to the hint," "as stated," or "as informed." You should treat these phenomena as suspicious cues that you naturally discovered through careful observation, and then carry out independent, objective, and professional analysis on this basis.

The sampled video frames are as follows

`{image_tokens}`

The anomaly types are as follows

- **Object Distortion**: detect deformation, mutation, clipping artifacts, or other structural defects in inanimate objects.
- **Human Distortion**: detect limb deformity, abnormal fingers, extra limbs, body splitting, or other structural defects in human subjects.
- **Motion Anomalies**: detect unnatural trajectories, abnormal jittering or flickering, and sudden disappearance or appearance of entities across frames.
- **Physical Violations**: detect unsupported floating, interpenetration between entities, or other behaviors that contradict real-world physics.
- **Character Anomalies**: detect unreadable text, abnormal text layout, or inconsistent textual content across frames.

Analysis steps and chain-of-thought requirements

In the `COT` field of the final output JSON object, strictly follow the following three-step structure to write the complete chain-of-thought analysis process, and ensure logical coherence (omission, merging, or skipping is forbidden):

Step 1: Analyze the visual content frame by frame, and determine whether issues such as object distortion, human distortion, motion anomalies, physical violations, and character anomalies exist.

Analyze each frame independently, starting from Frame 0. Each frame must begin with "Frame N:" (where N is the specific frame index). Please ensure that Step 1 contains a complete description of all frames. A complete paragraph must be written separately for Frame 0, Frame 1, Frame 2, . . . , Frame `{frame_count_index}` . Expressions such as "Frame 0–2" or "the first few frames" are forbidden. The following three aspects must be covered in order and in full. If an aspect does not apply, this must be explicitly stated and must not be omitted.

Analyze each frame independently, starting from Frame `{new_start}` . Each frame must begin with "Frame N:" (where N is the specific frame index). Please ensure that Step 1 contains a complete description of all frames. A complete paragraph must be written separately for Frame `{new_start}` , each subsequent frame, . . . , Frame `{new_end}` . Expressions such as "Frame 0–2" or "the first few frames" are forbidden. The following three aspects must be covered in order and in full. If an aspect does not apply, this must be explicitly stated and must not be omitted.

1. **Human analysis** (if there is no visible person, you must explicitly write "no visible person"), and the analysis must explicitly state whether an anomaly exists. Clothing: precisely describe the clothing type, color, clarity of patterns or textures, and all accessories. Posture: describe the overall pose and specifically describe the positions of the limbs. Expression: characterize eyebrow shape, gaze direction, and mouth state, and evaluate naturalness. Hand state: describe each finger one by one, whether the hand is holding an object, and hand details.
2. **Object analysis**, and the analysis must explicitly state whether an anomaly exists. Object identification: list all visible objects. Object attributes: describe the shape, color, and texture clarity of each object, and evaluate whether its size proportion is reasonable. Spatial relationship: state the positions between objects and people, and between objects and objects.
3. **Environment analysis**, and the analysis must explicitly state whether an anomaly exists. Scene: clearly specify the background type. Lighting: judge the direction of the main light source, and describe the position, length, and softness or hardness of shadows. Environment details: check whether the background contains blur, repeated textures, geometric distortion, pixel tearing, or unnatural patterns.

Figure S4: **Prompt for Video Anomaly Analysis CoT Creation (Part 1 of 2).** shared content; turn-one long-video-specific content; turn-two short-video-specific content; placeholders that will be replaced by specific inputs; conditional branches.

Prompt for Video Anomaly Analysis (Continued)

Step 2: Inter-frame coherence and motion trajectory analysis. Based on the human, object, and environment information in Step 1, determine whether anomalies exist.

Using information from multiple preceding and following frames, evaluate the following dimensions one by one. If an anomaly exists, give the corresponding description; if there is no change or no anomaly, you must still explicitly state "no anomaly": whether there are any issues related to object distortion, human distortion, motion anomalies, physical violations, or character anomalies.

Step 3: Overall video anomaly assessment

Based on all details in Step 1 and Step 2, comprehensively determine whether anomalies exist in the following five problem types. If an anomaly exists, the reason must be given; if no anomaly exists, write "no anomaly": Object Distortion, Human Distortion, Motion Anomalies, Physical Violations, Character Anomalies.

Output format requirements

You must output one and only one valid JSON object. Do not wrap it in a Markdown code block, and do not output any extra text. The JSON structure is as follows:

```
{
"COT": "Write the complete Step 1 / Step 2 / Step 3 chain-of-thought analysis process here.",
"status": "abnormal / normal",
"anomalies": [
{
"Attributed Time Region": "Fill in the frame range.",
"Attributed Label": "Fill in the specific anomaly category.",
"Reason for Anomaly": "Concrete anomaly description with the object name and the specific phenomenon.",
"Problem Region": "Semantic description of the problematic location.",
"BBOX": {
  "Frame 0": [xmin, ymin, xmax, ymax],
  "Frame 1": [xmin, ymin, xmax, ymax]
}
}
]
}
```

Strict rules for filling the `anomalies` field

- When `status` is "normal": this means that the visual content has no defects, and `anomalies` must be an empty list `[]`. Do not fabricate descriptions of normal content in the list.
- When `status` is "abnormal" and there is only one problem: output one dictionary with complete fields in the `anomalies` list.
- When `status` is "abnormal" and there are multiple problems: output multiple dictionaries in the `anomalies` list, separating different anomalies so that each dictionary corresponds to one independent anomaly point.

Regarding the BBOX coordinate output format: the key is the re-indexed frame number, such as `"Frame 0"` and `"Frame 1"`; the value is an integer coordinate array `[xmin, ymin, xmax, ymax]`; the BBOX dictionary must cover all frames in the snippet and must not omit any frame.

Please begin your analysis.

Figure S5: **Prompt for Video Anomaly Analysis CoT Creation (Part 2 of 2).** White indicates shared content, blue indicates long-video-specific content, orange indicates short-video-specific content, green indicates placeholders, and pink indicates conditional branches.

Prompt for Consistency Verification

Abnormal

You are an expert reviewer for video quality assessment. Please strictly judge whether the following AI-generated video analysis report reasonably identifies and explains the human-annotated errors.

[Human-Annotated Error Information]

Segment ID: `{segment_id}`

Video ID: `{video_id}`

Annotation conclusion: the video has problems

Specific errors:

`{gt_error_str}`

[Complete AI Analysis Report (JSON)]

`{full_analysis_str}`

Absolute strict evaluation criteria

The judgment can be "consistent" only if **all** of the following conditions are satisfied:

1. **The error type is exactly the same**: the error category used and described by the AI model must be completely consistent with the human annotation.
2. **The erroneous object is essentially the same**: the erroneous object identified by the AI model must refer to the same entity as the human annotation. the erroneous object identified by the AI model must refer to the same entity or the same region as the human annotation.
3. **The error phenomenon is essentially the same**: the error phenomenon described by the AI model must be identical in underlying nature to the phenomenon in the human annotation.
4. **The time range matches strictly**: the time range identified by the AI model must match the time range in the human annotation under the following criteria: the start-frame deviation must be no more than 2 frames; the end-frame deviation must be no more than 2 frames; the overall overlap of the time range must be at least 85%. the time range identified by the AI model must be broadly consistent with the time range in the human annotation, and both the start-frame deviation and the end-frame deviation must not exceed 2 frames.
5. **There is an exact one-to-one correspondence**: each analysis item produced by the AI model must form an exact one-to-one match with each human-annotated error description.

Please output only "consistent" or "inconsistent," and do not add any other explanation.

Figure S6: **Prompt for Consistency Verification.** Colors follow Fig. S4.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions, scope, and empirical claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E discusses the evaluation scope, missing release artifacts, absent multi-seed confidence intervals, and remaining generalization gaps.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not claim formal theorems or proofs; it presents model objectives and implementation equations, with additional derivations collected in Appendix C.1.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sections 3–4 and Appendices B.1–C.3 describe the dataset construction, training pipeline, GRPO formulation, prompt templates, benchmark setup, and inference configuration used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The dataset and generated-video assets are too large to be packaged directly in the supplementary material; at the anonymized submission stage, the paper and appendix instead document the methodology and experimental procedure in detail.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 specifies the backbone, optimization stages, learning rates, LoRA settings, and benchmark construction, while Appendices B.1–C.3 provide the supporting implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiments already require large-scale training runs at the 16-GPU scale, making multi-seed repetitions, confidence intervals, and statistical significance tests difficult to conduct within the available compute budget.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix C.2 reports GPU counts, training steps, wall-clock times, batch sizes, and optimization settings for the CaC training stages, while Appendix C.4 reports the 32-GPU Flow-GRPO alignment setup. The specific GPU models are not disclosed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work focuses on generated-video analysis rather than personal user data and discusses annotation and misuse considerations in Appendix E and Appendix B.1.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix E discusses both the benefits of improved anomaly detection and the risks of enabling stronger synthetic-video generation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The submission does not release high-misuse-risk or dual-use generative models or datasets; potential misuse considerations are nevertheless discussed in Appendix E.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: External models, datasets, and evaluation tools used in the study are credited through citations in the main text and appendix, and are used only under their intended research/evaluation settings; the submission does not redistribute third-party assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Section 3.1 and Appendices B.1–B.3 document the CaC Dataset and CaC-Bench, including annotation protocol, prompt synthesis, and dataset statistics.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Appendix B.1 summarizes the annotation workflow. The annotators were compensated for their work, but institution-level confidentiality requirements prevent us from disclosing detailed compensation information in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve an interventional human-subject study or collection of personal/sensitive participant data; human effort is limited to expert annotation of generated videos.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 3.1 and Appendix B.2 describe the use of Gemini-2.5-Pro for CoT synthesis during dataset construction.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.