

Appendix

Content of Appendices

Section A. The ViF Dataset.

- § A.1. Definition of Each Artifact Category.
- § A.2. Annotation Platform.
- § A.3. Chain-of-Thought Annotation Prompt Design.
- § A.4. Annotation Details, Dataset Statistics, and

Training Settings.

- § A.5. Detailed Statistics of ViF-CoT-4K.
- § A.6. Generated Video Examples.

Section B. Analysis of Generalization Ablations.

Section C. Analysis of Artifact Cues.

Section D. Analysis of Baselines' Detection Capabilities.

- § D.1. Binary Detectors.
- § D.2. Off-the-Shelf MLLMs.
- § D.3. Existing MLLM-based Detectors.

Section E. Additional Examples.

- § E.1. Design of Prompts.
- § E.2. Examples of Skyra's Responses.

Section F. Broader Impacts.

Section G. License.

A. The ViF Dataset

A.1. Definition of Each Artifact Category

We provide detailed definitions of each category in our artifact taxonomy (Section 3.2) as follows.

Low-Level Forgery. This group summarizes characteristic visual cues that frequently make current AI-generated videos appear “unnatural”. These cues typically do not explicitly violate physical laws, but reflect systematic limitations of mainstream video generation models.

- **Texture Anomaly.** This category focuses on abnormal patterns in local textures.
 - *Structure Anomaly.* Regions with rich structures (e.g., fences, grids, lattices) exhibit unnatural distortion, twisting, or interlacing, leading to inconsistent or implausible geometric patterns.
 - *Texture Jittering.* Surface textures show high-frequency flickering or drifting over time, manifesting as crawling patterns, grid-like noise, or temporally unstable blur, instead of stable, physically plausible textures.
 - *Unnatural Blur.* Blur and degradation patterns differ from typical natural degradations, such as Gaussian blur or compression artifacts. The blur may be spatially inconsistent, texture-selective, or temporally unstable in a way rarely observed in real videos.
- **Color and Lighting Anomaly.** This category captures implausible color or illumination patterns that deviate from natural imaging conditions.
 - *Color Over-saturation.* Certain regions exhibit excessively saturated or overly vivid colors (often in blue, red,

or green), with insufficient tonal variation or shading, making the area visually stand out unnaturally.

- *Lighting Inconsistency.* Global or local illumination changes abruptly or violently over time, or shows strong intensity fluctuations that cannot be explained by realistic changes of light sources, exposure, or scene configuration.
- **Motion Forgery.** This category describes unnatural camera-related motion patterns.
 - *Camera Motion Inconsistency.* The apparent camera motion is abnormal, such as erratic zooming in/out, unnatural high-frequency panning, or irregular shaking. These artifacts are often accompanied by inconsistent changes in object positions, scales, or spatial relations that do not match a physically plausible camera trajectory.

Violation of Laws. This group contains artifacts that clearly violate real-world constraints, including object permanence, physical laws, causality, and basic common sense. Detecting these cues generally requires spatio-temporal reasoning and background knowledge about how objects and scenes behave in reality.

- **Object Inconsistency.** This category focuses on violations of object permanence and identity over time.
 - *Abnormal Object Disappearance.* An object disappears suddenly during its motion without any plausible interaction or occlusion. For example, a runner on a track vanishes abruptly while continuing to move forward.
 - *Abnormal Object Appearance.* An object suddenly appears and starts to move without any reasonable cause or prior indication. For example, a new runner appears out of nowhere on the track in the middle of the video.
 - *Person Identity Inconsistency.* The identity of a person changes over time, especially in facial features or other stable identity cues. For example, a person's face disappears and reappears with clearly different facial characteristics, leading to a mismatch in perceived identity.
 - *General Object Identity Inconsistency.* The identity of a generic object changes significantly over time without any obvious external cause. For example, a chair being rotated by a person ends up with a drastically different color or structure compared to its initial state.
 - *Shape Distortion.* Rigid objects exhibit non-rigid deformations during motion. For instance, a human body suddenly scales up and down or undergoes frequent surface distortions and twisting that are incompatible with rigid-body motion.
- **Interaction Inconsistency.** This category captures physically implausible interactions between multiple objects.
 - *Abnormal Rigid-Body Crossing.* Rigid objects that should remain non-interpenetrating instead of intersecting or passing through each other. For example, a bar-

bell that should move in front of a person’s body passes unrealistically through the person’s head.

- *Abnormal Multi-Object Merging*. Two or more distinct objects gradually or abruptly merge into a single object during motion, without any plausible explanation (e.g., three people in motion merge into two).
- *Abnormal Object Splitting*. A single object splits into multiple distinct objects during motion, again without any reasonable cause (e.g., one person splits into two separate people).
- *General Interaction Anomaly*. Other abnormal or implausible phenomena occurring during interactions between two or more objects, such as missing collisions, inconsistent contact, or contradictory occlusion relations.
- **Unnatural Movement**. This category denotes motion patterns that contradict the typical kinematics of humans, animals, or objects.
 - *Unnatural Human Movement*. Human body motion deviates from normal biomechanics or everyday experience. For example, a person walks without leg crossing, exhibiting pure lateral sliding of the legs instead of realistic gait cycles.
 - *Unnatural Animal Movement*. Animal motion is incompatible with known locomotion patterns. For example, a running horse moves its hind legs in parallel translation without proper alternating strides.
 - *Unnatural General Object Movement*. Objects other than humans and animals follow trajectories or undergo transformations that are inconsistent with real-world dynamics, such as erratic acceleration, unnatural smoothness, or implausible temporal discontinuities.
- **Violation of Causality Law**. This category collects artifacts that violate physical laws or general causal relationships.
 - *Violation of Physical Laws*. The motion of objects contradicts basic physical principles, such as force–acceleration relationships or conservation laws. For example, a ball moves or changes velocity in the absence of any visible force, or instantaneously teleports at unrealistic speeds.
 - *Violation of General Causality Violation*. Events occur without observable causes, or actions fail to produce their expected effects. For example, a boy spills milk onto a table, but no milk traces appear on the table surface.
- **Violation of Common Sense**. This category covers structural or semantic inconsistencies that conflict with basic commonsense knowledge.
 - *Abnormal Human Body Structure*. The generated human body deviates from normal anatomical structure. Examples include extra or missing body parts (e.g., two heads, three or six fingers), or impossible body bending

that is incompatible with human physiology.

- *Abnormal General Object Structure*. Non-human objects exhibit structures that are inconsistent with their typical shapes or assembly, such as missing essential components or impossible connections.
- *Text Distortion*. Text appearing in the scene is severely distorted, malformed, or rendered as illegible gibberish without coherent semantic content, beyond mild degradation commonly observed in real footage.

A.2. Annotation Platform

Our annotation platform presents each AI-generated video alongside its corresponding real counterpart in a synchronized comparison view (Figure 7). This side-by-side layout allows annotators to directly contrast suspicious regions in the fake video with how the same scene should plausibly appear in real footage, making it easier to distinguish genuine physical phenomena from artifacts that only occur in AIGC videos. For every identified clue, annotators are required to select a fine-grained artifact category, provide detailed textual explanations for both the fake and real videos, and supply precise spatio-temporal annotations by marking time spans and bounding boxes in both streams. By enforcing mirrored annotations on fake–real pairs, the platform encourages annotators to explicitly encode both “what is wrong” in the generated video and “what is normal” in the real video, guiding the model toward learning an unbiased perceptual representation that treats real and synthetic content in a symmetric manner.

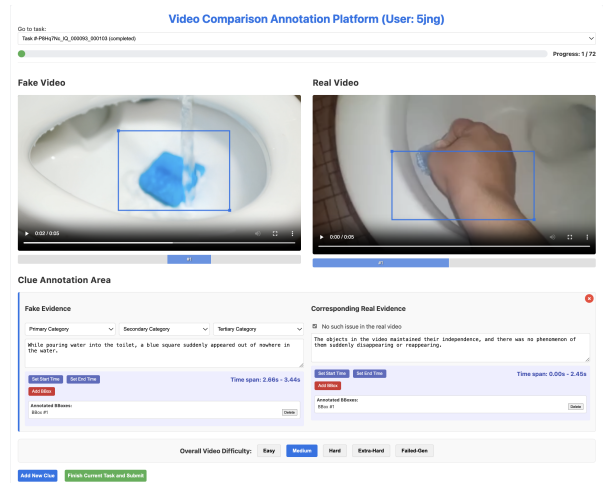


Figure 7. Annotation platform UI.

A.3. Chain-of-Thought Annotation Prompt Design

To transform concise human annotations into training-ready step-by-step supervision, we design a structured prompt for Gemini-2.5-Pro that operates on each fake–real video pair. For every annotated instance, the model receives sampled frames from the fake and real videos to

gether with the curated artifact **Type**, **Textual Explanation**, **Timestamps**, and **Bounding Boxes**, and is instructed to produce two independent CoT strings: one that carefully discovers all artifacts in the fake video and one that systematically clears the corresponding regions in the real video. The prompt enforces a standardized JSON output format (with separate `fake_cot_annotation` and `real_cot_annotation` fields), requires explicit tagging of temporal spans and spatial regions, and guides the reasoning process through an *observe-understand-draft-review-conclude* workflow with in-context examples. This design allows us to automatically expand precise but terse human labels into rich, consistent CoT supervision suitable for SFT. The complete prompt is provided in Figure 13

A.4. Annotation Details, Dataset Statistics, and Training Settings

We provide a comprehensive summary of dataset and training specifications in Table 5.

Annotation Protocol. A total of 25 trained professional annotators participated in the annotation process. Each annotator followed detailed guidelines covering artifact definitions (Section A.1), annotation platform usage (Section A.2), and quality standards. The annotation workflow proceeded as follows: (1) annotators watched each AI-generated video alongside its real counterpart in the synchronized comparison view (Figure 7); (2) for each identified artifact, they selected a fine-grained category from our taxonomy, provided textual explanations, and marked precise temporal spans and bounding boxes; (3) annotations underwent multiple review cycles to ensure quality and inter-annotator consistency.

Video Filtering. To ensure the quality and relevance of the collected videos, we apply a two-stage filtering pipeline: (i) *automatic semantic consistency checking* via GPT-4o-mini [46] between the AI-generated and real videos, which removes pairs with significant semantic drift; and (ii) *manual filtering* during annotation, where annotators remove static, ambiguous, or severely distorted samples that are unsuitable for meaningful artifact analysis.

A.5. Detailed Statistics of ViF-CoT-4K

We further reveal the statistic details of ViF-CoT-4K and ViF-Bench, including a detailed report of the proportion of different types of artifacts annotated in ViF-CoT-4K (Table. 9), and technical details of the video generation model in ViF-CoT-4K and ViF-Bench (Table. 10).

A.6. Generated Video Examples

We demonstrate the quality of our dataset and benchmark by showing several examples randomly selected from the ViF-Bench (Figures 14& 15). As shown in the image, fake samples in our dataset are generated by latest video generation

Table 5. Dataset and training specifications.

Dataset Statistics	
Video duration	5 seconds (real & fake, train & test)
Total frames per video	81 frames
Resolution	256p (short side)
Training Settings	
Sampled frames	16 frames per video (uniformly sampled)
Input resolution	256p (short side)
SFT learning rate	1e-5
SFT epochs	5
RL actor learning rate	5e-7
RL KL coefficient	0.02
Batch size per device	1
Hardware	8× NVIDIA H200 GPUs

Table 6. Generalization ablation study across in-generator, cross-generator, and cross-dataset settings on ViF-Bench and GenVideo.

Setting	In-Generators			Cross-Generators			Cross-Dataset		
	Acc	R	F1	Acc	R	F1	Acc	R	F1
w/o CoT	57.27	15.67	26.94	53.99	9.29	16.55	51.89	6.78	14.66
w/o SFT	70.91	87.27	75.00	75.87	98.82	80.36	<u>68.34</u>	<u>34.50</u>	<u>46.00</u>
Skyra-SFT	<u>92.62</u>	<u>89.49</u>	<u>92.13</u>	<u>88.95</u>	82.42	<u>87.20</u>	63.98	28.70	41.00
Skyra-RL	93.58	93.22	93.45	89.84	<u>86.10</u>	88.80	71.78	45.60	59.00

models, and are closely aligned with their real counterparts to mitigate their gap in semantics and format.

B. Analysis of Generalization Ablations.

Generalization Analysis: RL is the primary driver of cross-domain generalization. We further isolate each component’s contribution to generalization by evaluating under in-generator, cross-generator, and cross-dataset settings (Table 6). While SFT significantly improves in-domain accuracy, it also introduces mild overfitting, as cross-dataset accuracy drops from 68.34% to 63.98%. The subsequent RL stage recovers and extends generalization, yielding a +7.8% cross-dataset accuracy gain over SFT while preserving in-domain performance. This suggests that RL encourages the model to discover intrinsic artifact cues rather than generator-specific patterns.

C. Analysis of Artifact Cues

To provide insight into what visual and temporal cues Skyra relies on for detection, we analyze the distribution of artifacts detected by Skyra on ViF-Bench, as shown in Table 7.

Among all detected artifacts, 82.8% belong to *Violation of Laws* while only 17.2% are *Low-Level Forgery*, indicating that Skyra primarily performs semantic reasoning about physical and logical inconsistencies rather than relying on low-level generator fingerprints. Within the *Violation of Laws* category, *Object Inconsistency* accounts for 28.1%, with *Shape Distortion* (15.2%) being the single most dominant artifact type. This reflects a shared difficulty across cur-

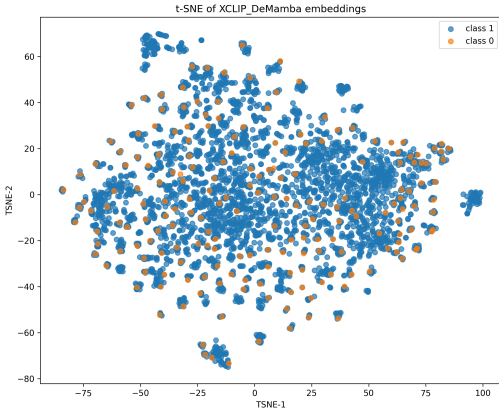


Figure 8. The T-SNE result of Demamba.

rent video generators in maintaining rigid-body constraints during dynamic scenes. *Interaction Inconsistency* (10.0%) and the remaining categories (44.7%) further demonstrate that Skyra attends to diverse physical violations spanning object permanence, causality, and commonsense.

The relatively low proportion of *Low-Level Forgery* (17.2%) suggests that Skyra does not overfit to superficial texture or color cues, which tend to be generator-specific and less generalizable. Instead, its reliance on semantic-level violations aligns with how humans identify AI-generated content, contributing to its stronger cross-generator and cross-dataset generalization observed in Table 6.

Table 7. Distribution of artifacts detected by Skyra on ViF-Bench, organized by our hierarchical taxonomy.

L1 Category (Ratio)	L2 Category (Ratio)	L3 Category	Ratio	
Low-Level Forgery (17.2%)	Texture Anomaly (11.2%)	Texture Jittering	3.1%	
		Structure Anomaly	3.5%	
	Others (6.0%)	Others	4.6%	
		
Violation of Laws (82.8%)	Object Inconsistency (28.1%)	Shape Distortion	15.2%	
		Abnormal Object Appearance	5.8%	
		Others	7.1%	
	Interaction Inconsistency (10.0%)	Abnormal Multi-Object Merging	2.7%	
		General Interaction Anomaly	3.2%	
		Others	4.1%	
		Others (44.7%)

D. Analysis of Detection Capabilities

D.1. Binary Detectors

We take Demamba [5] and NSG-VD [84] as examples of underperforming classifiers that exhibit a strong tendency to label samples as fake.

Demamba: Through T-SNE visualization (Figure 8) and CAM heatmap analysis (Figure 9), we observe that in the T-SNE embedding space, fake and real samples are highly overlapping and difficult to separate. CAM heatmaps further reveal that, for real-labeled videos, the model consistently focuses on similar spatial locations across different samples,

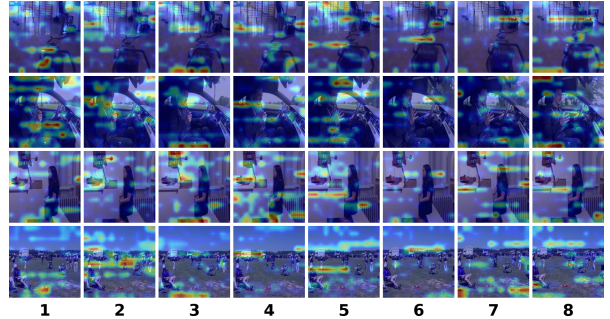


Figure 9. Visualization of Class Activation Maps (CAMs) produced by DeMamba on real video samples.

particularly in the third and fourth frames of the sequences. This suggests that the model may be overly sensitive to fixed visual patterns or preferred spatial locations in the scene, rather than learning generalized content-based cues such as human motion or manipulation traces. Overall, the model does not attend to regions that are discriminative for authenticity, but rather to textures outside the main content of the frame, indicating limited generalization.

NSG-VD: This classifier leverages a reference dataset, providing NSG feature baselines from real videos during inference. Specifically, the normalized spatiotemporal gradient (NSG) features of the reference data are used to model the distribution of real video dynamics, and the maximum mean discrepancy (MMD) between test videos and this reference distribution is computed. A test video is classified as AI-generated if its MMD exceeds a predefined threshold.

In our reproduction, the model achieved near-perfect AU-ROC on the validation set but behaved poorly on the test set, indicating a strong tendency to overfit. This overfitting may stem from the intrinsic sensitivity of NSG-based methods to subtle distributional shifts: the model struggles to generalize when the spatiotemporal dynamics of real videos deviate even slightly from those observed during training. When evaluating NSG-VD on the OOD GenVideo Benchmark, we consider it inappropriate to use the real samples within the GenVideo Benchmark as the reference dataset. Doing so would contradict the purpose of forgery detection and could introduce data leakage. Therefore, we retain the same reference dataset used during training. Similarly, in our robustness study, we do not use real samples under various degraded formats as reference data, because assuming access to degradation-specific real videos for each testing condition is impractical in real-world applications. This decision to reuse the training-time reference dataset may partly explain the suboptimal performance of NSG-VD in both experiments.

Overall, while the reference-dataset mechanism in NSG-VD provides strong detection capability when the distribution of the target data is known or partially accessible, its reliance on such reference information limits its effectiveness in OOD or noisy scenarios, where the real samples to

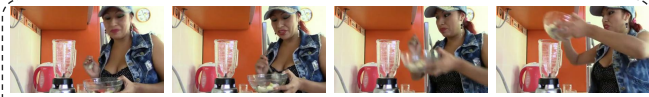


Artifacts Type: Abnormal Human Body Structure

GPT-4.1-mini

<think>The video frames depict a scene at an outdoor event with two men prominently in the foreground, one wearing a white shirt and hat and the other in a black shirt. The background shows a crowd, flags, and vehicles, including a police car. The sequence shows the two men interacting, with the man in white seemingly moving his arm toward the man in black. Step-by-step analysis: 1. From T=0.00s to T=1.31s, the man in white is holding a sign and then begins to move his arm. The motion blur on the man in white's arm and face is quite strong, especially at T=1.00s and T=1.31s, which is consistent with fast movement captured by a camera. This blur looks natural and typical of motion blur in real videos. 2. From T=1.62s to T=2.31s, the man in white's arm moves closer to the man in black, and the motion blur continues. **The interaction between the two men appears fluid, with no sudden jumps or unnatural transitions in their positions or arm movements.** 3. From T=2.62s to T=3.62s, the man in black turns his head toward the man in white, and the man in white's arm is near the man in black's chest. The motion blur on the man in white's arm and the slight blur on the man in black's face are consistent with natural movement and camera capture. 4. From T=4.00s to T=5.00s, the man in white is no longer in the frame, and the man in black is looking to his left. The background remains consistent, with no sudden changes in lighting or object appearance. 5. Throughout the sequence, the lighting on the two men and the background is consistent with a sunny outdoor environment. Shadows and highlights correspond well with the apparent sun position. 6. No signs of texture anomalies, unnatural blurring beyond expected motion blur, or lighting inconsistencies are observed. 7. **The interaction between the two men, including their relative positions and movements, appears physically plausible and continuous.** 8. No abnormal object appearances or disappearances are detected. The police car and background elements remain consistent. 9. The motion blur and slight focus changes are typical of real handheld or moving camera footage. Based on the above observations, there are no detectable artifacts or inconsistencies that would suggest manipulation or forgery. No artifacts to tag.</think>
<answer>Real</answer>

(a) False Negative on Fake Videos



Real Video

Gemini-2.5-flash

<think>The video frames show a woman in a kitchen, wearing a denim vest and a cap, preparing ingredients in a bowl and then attempting to pour them into a blender. Here's a step-by-step analysis:
1. **Initial Observation (T=0.00s - T=4.62s):** The woman is holding a clear glass bowl with chopped ingredients and a spoon, stirring them. A blender is on the counter to her left. The movements appear natural, and the lighting and textures seem consistent.
2. **Action Analysis (T=4.62s):** The woman begins to lift the bowl, preparing to pour its contents into the blender. Her right hand is holding the bottom of the bowl, and her left hand is holding the side.
3. **Critical Frame Analysis (T=5.00s):** At T=5.00s, the woman has lifted the bowl and is tilting it towards the blender. **However, a significant anomaly is observed.** The contents of the bowl, which were clearly visible as chopped ingredients (fruits/vegetables) in previous frames, are now completely gone. **The bowl is empty, yet her hands are positioned as if she is pouring something.** This is a clear **Abnormal Object Disappearance** in <[4.62s, 5.00s]> at <[600, 400, 800, 600]> (referring to the contents of the bowl). The bowl itself is still present, but its contents have vanished.
4. **Further Examination of T=5.00s:** The woman's facial expression and body posture suggest she is actively pouring, but the bowl is empty. This creates a **Violation of General Causality Law** in <[5.00s, 5.00s]> at <[0, 0, 1000, 1000]> as the action of pouring implies contents, which are absent.
Based on the sudden and complete disappearance of the bowl's contents while the pouring action is being performed, the video exhibits a clear manipulation.</think>
<answer>Fake</answer>

(b) False Positive on Real Videos

Figure 10. Response examples of off-the-shelf MLLMs.

be evaluated are either unavailable or should not be incorporated into the reference set in the first place.

D.2. Off-the-Shelf MLLMs

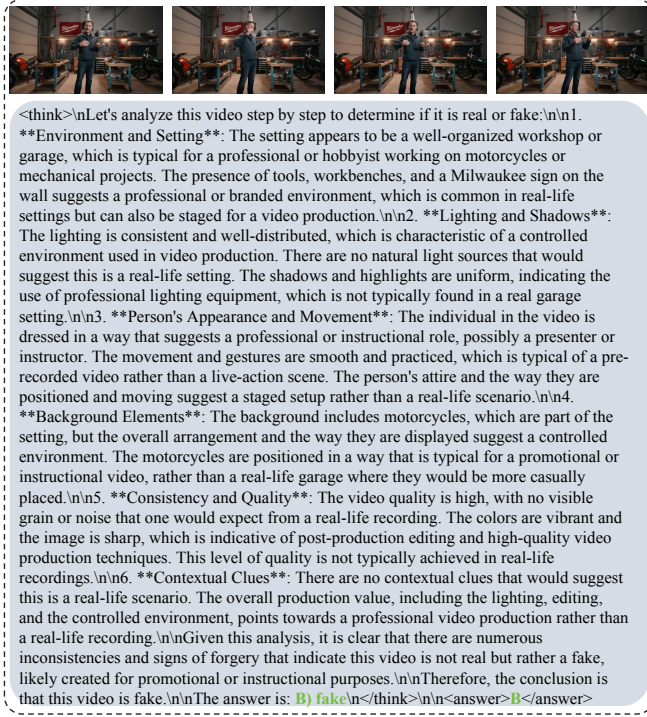
Figure 10 provides concrete examples of the failure modes discussed in the main paper when directly prompting off-the-shelf MLLMs for AI-generated video detection. In panel (a), we show *false negatives* on fake videos. Even with explicit chain-of-thought instructions, most models focus on high-level semantics and overall visual appeal (e.g., “the scene looks natural” or “the movements are smooth”) while overlooking intrinsic forgery cues such as inconsistent geometry or physics-violating motion. As a result, they confidently classify clearly synthetic videos as real and provide rationales that largely describe the content instead of analyzing subtle spatiotemporal artifacts.

Panel (b) shows the opposite pattern. Models such as Gemini-2.5-flash [11] tend to over-interpret natural video degradations, including compression artifacts, motion blur, and low-light noise, as evidence of forgery. In these cases, the model produces detailed yet incorrect explanations that attribute the degradations to “AI generation” rather than common acquisition or post-processing effects. This confirms our quantitative findings that off-the-shelf MLLMs tend to conflate quality with authenticity: they are sensitive to superficial visual cues but struggle to distinguish genuine forgery artifacts from benign imperfections in real-world videos.

D.3. Existing MLLM-based Detectors

We further analyze BusterX++ [72, 73], a recent MLLM-based detector that adapts pretrained models for AIGC video detection. Figure 11 (a) shows a *success case* where BusterX++ correctly identifies an AI-generated video. In such scenarios, the synthetic content exhibits obvious stylistic or aesthetic discrepancies from typical real videos (e.g., overly smooth textures or globally inconsistent lighting), which align well with the model’s training biases and allow it to reach the correct decision.

However, panels (b) highlight the limitations of relying primarily on global scene appearance. Here, the AI-generated clip contains subtle but critical physics-violating artifacts, which humans readily notice. BusterX++, however, focuses on the overall coherence and visual quality of the scene and fails to attend to these localized spatiotemporal inconsistencies, leading to an incorrect “real” prediction. Together, these examples corroborate our main observation that current MLLM-based detectors behave more like general content describers: they emphasize superficial, distribution-level cues and natural degradations, but are not yet equipped to systematically discover and reason about intrinsic forgery artifacts that are crucial for reliable AI-generated video detection.



(a) Success Cases



(b) Failed Cases

Figure 11. Response examples of existing MLLM-based detector, BusterX++ [73].

E. Additional Examples

E.1. Design of Prompts

We specify the system and user prompt that Skyra uses in Figure 12. The system prompt specifies the model's role as an AI video analyst, clearly defines the output format (a <think> reasoning block followed by a one-word <answer> verdict), and constrains the reasoning to our artifact taxonomy, requiring that all findings be tagged with explicit categories, time spans, and bounding boxes. In contrast, the user prompt focuses on supplying multimodal evidence: we interleave sampled frames with their timestamps (e.g., "[T=0.00s] <image>" ... "[T=5.00s] <image>"), so that the model can reason over the evolution of the scene, align artifacts with precise temporal positions, and improve its ability to detect subtle, time-dependent inconsistencies.

E.2. Examples of the Responses of Skyra

We provide the inference examples of Skyra on more samples in ViF-Bench. Figures 16 and 15 demonstrate its responses when encountering real videos. Figure 18- 25 exhibit different types of evidence that Skyra uses when determining that a video is AI-generated.

F. Broader Impacts

Our work is motivated by the growing societal risks posed by AI-generated videos, including large-scale misinformation,

System Prompt:

You are an expert AI video analyst. Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.

You MUST output your entire analysis using the following structure:

- A <think>...</think> block containing your detailed reasoning.
- An <answer>...</answer> block containing the final, one-word verdict: 'Fake' or 'Real'.

Inside the <think> block, you MUST:

- Start by briefly describing the overall content of the video frames.
- Follow a detailed, step-by-step "discovery" or "verification" process.
- When you identify an artifact (or clear a region), you MUST use a valid L3 Category Name from the "Artifact Category Definitions" provided below.
- You MUST embed your finding using the following exact tag structure:
<type>L3 Category Name</type> in <time>[startTime, endTime]</time> at <bbox>[x1, y1, x2, y2]</bbox>
- If multiple artifacts are present, you must find and tag all of them in temporal order.
- Your entire reasoning process must be self-contained

...

Artifact Category Definitions (Valid L3 Categories for the <type> tag)

...

[Artifact Taxonomy Here]

User Prompt:

Here are the video frames and their corresponding timestamps:

[T=0.00s] <image>

...

[T=4.67s] <image>

[T=5.00s] <image>

Please analyze the video frames, determine if the video is real or fake, and provide your reasoning

Figure 12. System prompt and user prompt design.

impersonation, and erosion of trust in authentic media. By focusing on interpretable, artifact-centric detection, Skyra aims to provide not only predictions but also grounded visual evidence that can assist journalists, fact-checkers, regulators,

Table 8. License of source datasets in ViF-CoT-4K and ViF-Bench.

Dataset	License
Kinetics-400 [29]	CC BY 4.0
Panda-70M [6]	Snap Inc. Non-Commercial Research
HD-VILA-100M [78]	AGPL-3.0

and platform moderators in assessing the authenticity of suspicious content. The ViF-CoT-4K dataset and ViF-Bench further offer a standardized testbed for evaluating new detectors on diverse, up-to-date generators, which may contribute to more reliable and transparent AIGC safety tools.

At the same time, releasing a detailed artifact taxonomy, benchmark, and detector introduces dual-use concerns. In principle, insights into the failure modes of current detectors could inform future attempts to design more robust and evasive generative models. We believe that, on balance, the benefits of enabling the research community, civil society, and industry to build stronger and more interpretable defenses outweigh these risks. To mitigate potential misuse, our datasets contain only curated, non-sensitive content, and we emphasize that Skyra is intended to support human-in-the-loop verification rather than fully automated decision making or mass surveillance. We encourage downstream users to deploy our models and data in accordance with relevant regulations, to combine them with complementary safeguards such as provenance and watermarking, and to continuously stress-test detectors as the landscape of generative video models evolves.

G. License

ViF-CoT-4K and ViF-Bench are provided to the community under CC BY 4.0 license. By downloading our dataset from our website or other sources, the user agrees to adhere to the terms of CC BY 4.0 and the licenses of the source datasets. Licenses of the source datasets are listed in the Table 8.

You are an expert AI video analyst. Your task is to generate meticulous, step-by-step Chain-of-Thought (CoT) reasoning for training a deep learning model. You will be given a set of FAKE video frames, a set of corresponding REAL video frames, and a ground-truth JSON annotation.

Your goal is to generate *two independent annotations* based on this data.

1. Input You Will Receive

1. **[FAKE VIDEO FRAMES]**: 16 sampled frames from the fake video.
2. **[REAL VIDEO FRAMES]**: 16 sampled frames from the real video.
3. **[HUMAN ANNOTATION]**: A JSON object (with Chinese descriptions) that you must use as a reference for *what* to find (in the fake) or *what to clear* (in the real).

2. Output You MUST Generate

You must output *only* a single, valid JSON object with two keys: "fake_cot_annotation" and "real_cot_annotation".

```

...json
{
  "fake_cot_annotation": "...",
  "real_cot_annotation": "..."
}
...

```

3. Rules for "fake_cot_annotation"

The string must follow this structure:
`<think>...thinking process...</think>\n<answer>Fake</answer>`

The thinking process **MUST**:

1. Be **fully self-contained**. It must **ONLY** analyze the **[FAKE VIDEO FRAMES]**. Do NOT mention or allude to a real video.
2. Start by describing the video content, then actively scan for artifacts (e.g., "I'm checking the lighting...", "The background texture...").
3. **You MUST** find and integrate ALL pieces of evidence from the [HUMAN ANNOTATION] JSON in **temporal order** as a natural part of your discovery narrative (e.g., "First, I noticed...", "Then, as I kept watching...").
4. For each piece of evidence, embed its tags: `<type>L3 Category Name</type> in <[start, end]> at <[x1, y1, x2, y2]>`
5. Your reasoning must be in English. **Do NOT** repeat or quote the Chinese description. Use the Chinese text to **understand** what artifact to look for, then describe that artifact in your **own** English words.

4. Rules for "real_cot_annotation"

The string must follow this structure:
`<think>...thinking process...</think>\n<answer>Real</answer>`

The thinking process **MUST**:

1. Be **fully self-contained**. It must **ONLY** analyze the **[REAL VIDEO FRAMES]**. Do **NOT** mention "fake video," "counterpart," "hotspot," or any other data outside of the real frames.
2. Follow a "diligent inspection" narrative. Start by describing the video, then state you are conducting a thorough review for common generation artifacts.
3. **You MUST** proactively inspect and clear ALL regions corresponding to the [HUMAN ANNOTATION] evidence, in temporal order.
4. For each region inspected, reference the tags to specify **what** you are clearing: "...I am checking the area <[start, end]> at <[x1, y1, x2, y2]>."
5. After checking each tagged region, state in your **own** English words that it is clear (e.g., "the text here is stable," "the motion is fluid," "the objects interact correctly").
6. Conclude that the video is real.

5. Artifact Category Definitions

6. Examples (In-Context Learning)

Example 1 (Single Clue + Independent CoT)

[HUMAN ANNOTATION]:

```

...json
{
  "video_id": "b03HGh02Ak-2",
  "evidence": [
    {
      "fake": {
        "category": { "L1": "Violation of Laws", "L2": "Violation of Commonsense", "L3": "Text Distortion" },
        "description": "...",
        "temporal": { "start": 0.91, "end": 5.04 },
        "spatial": [ 0.4, 0.2, 0.6, 0.2 ]
      }
    }
  ]
}
...

```

YOUR REQUIRED OUTPUT:

```

...json
{
  "fake_cot_annotation": "<think>The video appears to be a news report with an anchor and on-screen graphics. I'm scanning the overall scene first. The anchor's face and movement seem stable, no obvious Shape Distortion. The lighting is consistent. My attention is drawn to the on-screen graphics. The main title graphic is sharp, but as I look closer at the text *below* the main title, something is wrong. The characters look like they are flickering and warping. Yes, upon inspection, this is a clear artifact. I see <type>Text Distortion</type> in <[0.91, 5.04]> at <[0.4, 0.2, 0.6, 0.2]> where the text is warped and has no clear semantic meaning.</think>\n<answer>Fake</answer>".
  "real_cot_annotation": "<think>I am analyzing this news report. I will conduct a thorough inspection for any signs of AI generation. The anchor's face and movement appear natural. The lighting is consistent. I am paying close attention to the on-screen text, as this is a common failure point. Specifically, I am examining the lower text area <[0.91, 5.04]> at <[0.4, 0.2, 0.6, 0.2]>. In these frames, the text is perfectly sharp, stable, and legible. It shows no signs of distortion. My review of this clip finds no artifacts.</think>\n<answer>Real</answer>".
}
...

```

Example 2 (Multiple Clues + Independent CoT)

[HUMAN ANNOTATION]:

```

...json
{
  "video_id": "2xgtqmQA0xg-0",
  "evidence": [
    { "fake": { "category": { "L3": "Shape Distortion" }, "temporal": { "start": 0, "end": 0.24 }, "spatial": [ 0.3, 0.6, 0.4, 0.8 ] } },
    { "fake": { "category": { "L3": "Abnormal Multi-Object Merging" }, "temporal": { "start": 0.79, "end": 1.22 }, "spatial": [ 0.1, 0.3, 0.3, 0.6 ] } },
    { "fake": { "category": { "L3": "General Object Identity Inconsistency" }, "temporal": { "start": 3.3, "end": 3.61 }, "spatial": [ 0.3, 0.4, 0.4, 0.6 ] } }
  ]
}
...

```

YOUR REQUIRED OUTPUT:

```

...json
{
  "fake_cot_annotation": "<think>This video shows a close-up of someone assembling small mechanical parts. I am observing the interaction between the hands and the components. The camera is steady. I'll watch the process chronologically. Right at the beginning, I'm focusing on the part in the lower-middle frame. Its shape seems... unstable. It's warping and deforming, which isn't right for a solid part. This is a clear <type>Shape Distortion</type> in <[0, 0.24]> at <[0.3, 0.6, 0.4, 0.8]>. I'll continue watching. Now, around the 1-second mark, the person is picking up another part. As the handheld part touches the fixed part, they... blend together? It looks like the two distinct objects are merging unnaturally. This is a <type>Abnormal Multi-Object Merging</type> in <[0.79, 1.22]> at <[0.1, 0.3, 0.3, 0.6]>. This is very suspicious. Let me keep watching. Later, around 3.3 seconds, my attention is drawn to the part in the person's hand. It was clearly gold-colored a moment ago, but suddenly it flashes and changes to black and silver, with no external cause. This is a <type>General Object Identity Inconsistency</type> in <[3.3, 3.61]> at <[0.3, 0.4, 0.4, 0.6]>. The video has multiple, severe artifacts.</think>\n<answer>Fake</answer>".
  "real_cot_annotation": "<think>I am carefully reviewing this assembly footage to verify its authenticity. I will check several key moments for common artifacts. First, I am checking the component at the start, in area <[0, 0.24]> at <[0.3, 0.6, 0.4, 0.8]>. The part is perfectly rigid and stable; its shape does not change. Next, I'm checking the part interaction around 1 second, in area <[0.79, 1.22]> at <[0.1, 0.3, 0.3, 0.6]>. The handheld part makes contact with the fixed part, but they remain distinct objects. The interaction is normal. Finally, I'm checking the object identity around 3.3 seconds, in area <[3.3, 3.61]> at <[0.3, 0.4, 0.4, 0.6]>. The part in the hand remains a consistent gold color throughout the clip. It does not change. This video passes all checks.</think>\n<answer>Real</answer>".
}
...

```

Figure 13. Chain-of-Thought Annotation Prompt.

Table 9. Hierarchical distribution of artifact categories (L1–L2–L3) in ViF-CoT-4K.

L1 Category	Ratio	L2 Category	Ratio	L3 Category	Ratio
Low-Level Forgery	17.2%	Texture Anomaly	11.2%	Structure Anomaly	3.5%
				Texture Jittering	3.1%
				Unnatural Blur	3.6%
		Color & Lighting Anomaly	5.4%	Color Over-Saturation	2.8%
				Lighting Inconsistency	2.5%
		Move Forgery	1.6%	Camera Motion Inconsistency	1.6%
Violation of Laws	82.8%	Object Inconsistency	28.1%	Abnormal Object Disappearance	3.6%
				Abnormal Object Appearance	5.8%
				Person Identity Inconsistency	1.1%
				General Object Identity Inconsistency	2.4%
				Shape Distortion	15.2%
		Interaction Inconsistency	10.0%	Abnormal Rigid-Body Crossing	2.7%
				Abnormal Multi-Object Merging	2.7%
				Abnormal Object Splitting	1.4%
				General Interaction Anomaly	3.2%
		Unnatural Movement	10.0%	Unnatural Human Movement	6.6%
				Unnatural Animal Movement	0.5%
				Unnatural General Object Movement	2.9%
		Violation of Causality Law	6.90%	Violation of Physical Law	4.1%
				Violation of General Causality Law	2.8%
		Violation of Commonsense	27.8%	Abnormal Human Body Structure	10.5%
Abnormal General Object Structure	3.2%				
Text Distortion	14.1%				

Table 10. Overview of video generation models used to synthesize forged samples in our dataset. “Ref. Cond.” denotes the typical conditioning modes (T2V: text-to-video, I2V: image-to-video, TI2V: text+image-to-video). For commercial closed-source systems, parameter sizes are not publicly disclosed and thus marked as “N/A (closed)”.

Model Name	Release Date	Parameter	Ref. Cond.	Inference	Sample Number	HyperLink
Wan2.1-1.3B-T	2025-02-25	1.3B	T2V	Local	750	Link
CogVideoX1.5-T	2024-08	5B	T2V	Local	744	Link
CogVideoX1.5-I	2024-08	5B	I2V	Local	760	Link
Wan2.2-TI2V-5B(T2V)	2025-08-28	5B (MoE)	T2V	Local	747	Link
Wan2.2-TI2V-5B(I2V)	2025-08-28	5B (MoE)	I2V	Local	748	Link
HunyuanVideo	2024-12-03	13B	T2V	Local	750	Link
HunyuanVideo-I2V	2025-05-06	13B	I2V	Local	968	Link
VACE-1.3B-T (Wan2.1-VACE-1.3B)	2025-05-14	1.3B	T2V	Local	165	Link
Wan2.2-T2V-A14B	2025-08-28	14B (MoE)	T2V	Local	165	Link
Wan2.2-I2V-A14B	2025-08-28	14B (MoE)	I2V	Local	165	Link
SkyReels-V2-T2V	2025-04-21	14B	T2V	Local	165	Link
SkyReels-V2-I2V	2025-04-21	14B	I2V	Local	164	Link
LTX-Video(T2V)	2025-05-06	13B	T2V	Local	165	Link
LTX-Video(I2V)	2025-05-06	13B	I2V	Local	165	Link
Gen4-Turbo	2025-04	N/A (closed)	I2V	API	121	Link
Hailuo-02	2025-06-18	N/A (closed)	T2V	API	137	Link
Pika-V2	2025-08-15	N/A (closed)	T2V	API	151	Link
Pixverse-V4-5	2025-05	N/A (closed)	T2V	API	152	Link
Kling-V1	2024-06	N/A (closed)	T2V	API	141	Link
Sora-2	2024-02-15	N/A (closed)	T2V	API	150	Link

Note. Dates are approximate and refer to the first public announcement or open release of the corresponding model family. For commercial systems with undisclosed architecture/size, “Parameter” is marked as “N/A (closed)”.

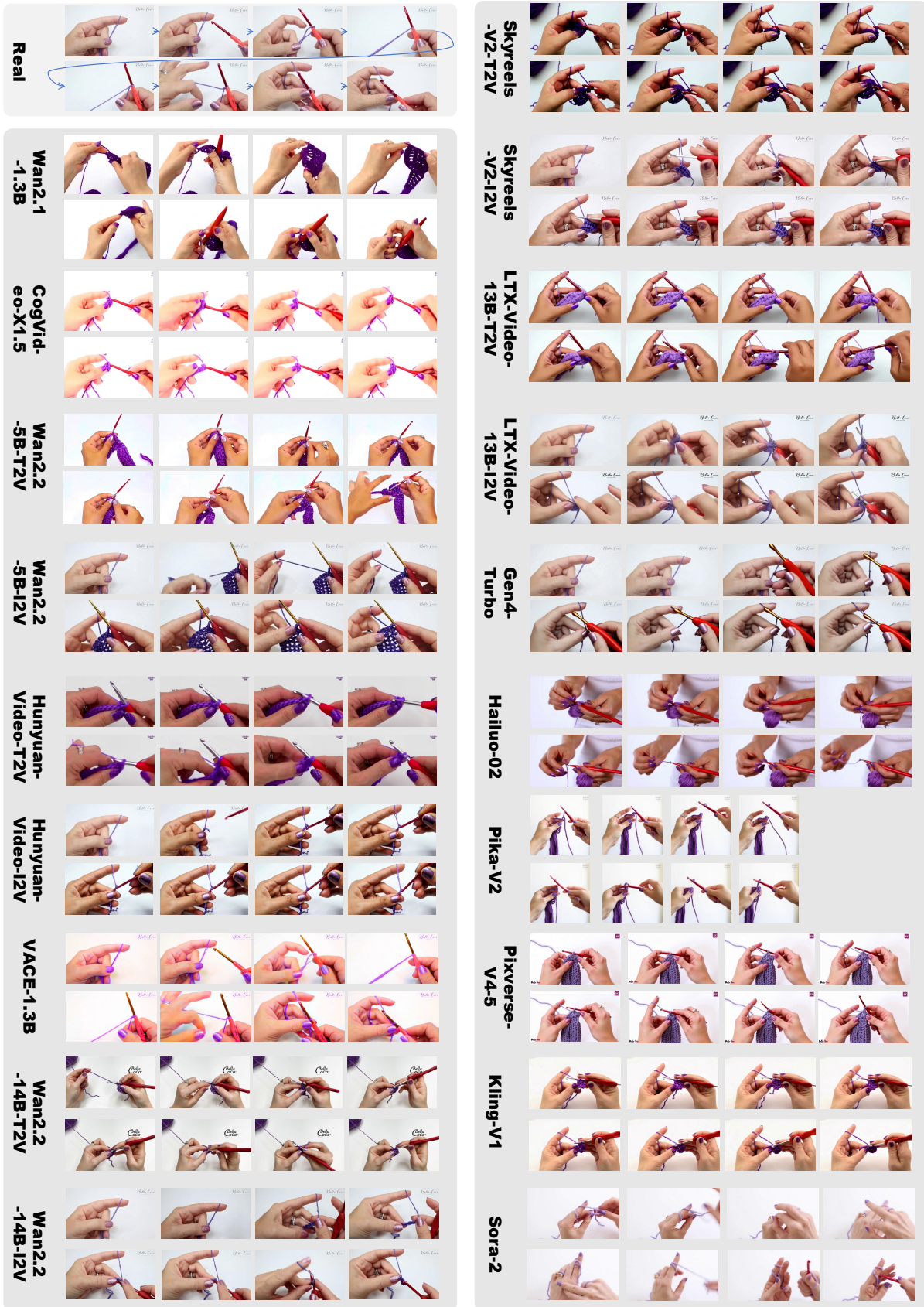


Figure 14. ViF-Bench Video Sample Examples-I

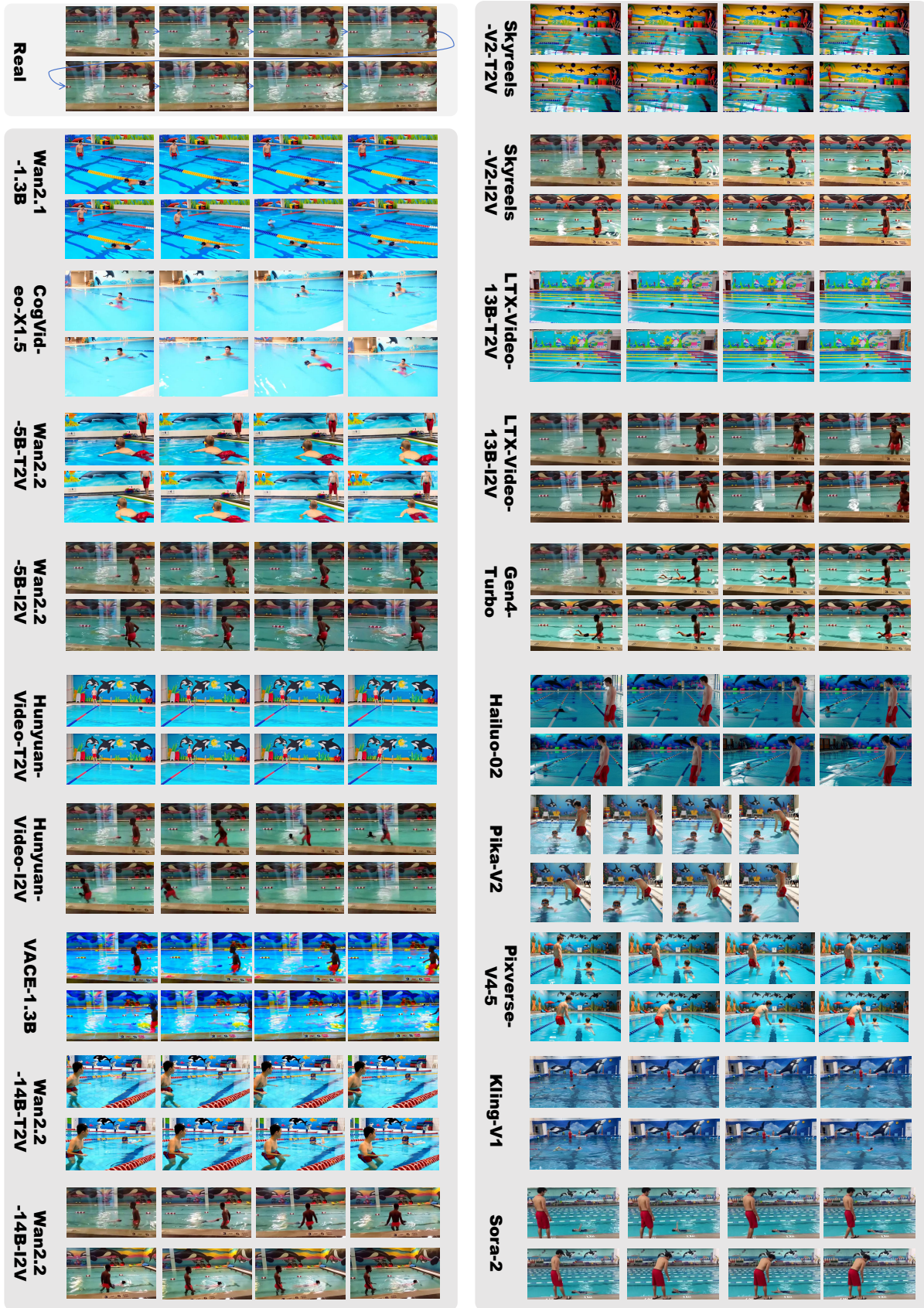


Figure 15. ViF-Bench Video Sample Examples-II

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>I am conducting a thorough analysis of this video, which shows a person preparing a pasta casserole in a glass dish. I will meticulously check for any signs of digital manipulation or generation artifacts. First, I am examining the initial action of the person placing a wooden spoon into the dish, specifically in the area $\langle t \rangle [0.0, 0.55] \langle /t \rangle$ at $\langle bbox \rangle [0.3, 0.2, 0.8, 0.6] \langle /bbox \rangle$. The spoon is held by the person and is clearly inserted by them into the pasta. The interaction is physically plausible and natural. Next, I am inspecting the movement of the spoon as it stirs the pasta, focusing on the region $\langle t \rangle [0.9, 1.55] \langle /t \rangle$ at $\langle bbox \rangle [0.3, 0.2, 0.8, 0.5] \langle /bbox \rangle$. The spoon moves fluidly and remains a solid, rigid object throughout the motion. There is no evidence of warping, bending, or any form of distortion. Finally, I am observing the person's hand from the side, in the area $\langle t \rangle [2.4, 3.06] \langle /t \rangle$ at $\langle bbox \rangle [0.1, 0.0, 0.4, 0.4] \langle /bbox \rangle$, to check for any anatomical inconsistencies. The hand and fingers maintain a normal, consistent shape and structure as they guide the spoon. After a detailed inspection, I find no artifacts or violations of physical laws. The video appears to be authentic.</think>
 <answer>Real</answer>

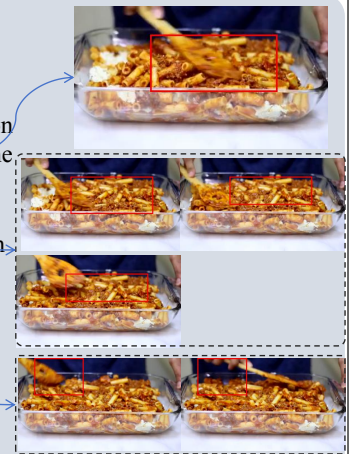


Figure 16. Skyra's Response Example on Real Videos, I

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>I am analyzing a video of a man with a distinctive white, curly wig speaking into a microphone. My goal is to conduct a thorough inspection to verify its authenticity. I will pay close attention to common areas where digital artifacts might appear, such as detailed textures. I am specifically examining the man's hair, as it is a complex texture that can be difficult to render correctly. I am checking the area $\langle t \rangle [0.0, 5.06] \langle /t \rangle$ at $\langle bbox \rangle [0.3, 0.1, 0.7, 0.8] \langle /bbox \rangle$ throughout the clip. Upon careful review, the hair's texture appears completely natural. I can see the definition of individual strands and curls, and the way light reflects off it is consistent with the environment. There is no unnatural blur, smudging, or static, grain-like noise. The texture is clear and moves realistically as the man speaks and gestures. Having found no evidence of digital manipulation, I conclude the video is authentic.</think>
 <answer>Real</answer>



Figure 17. Skyra's Response Example on Real Videos, II

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>The video shows a static shot of a city skyline across a body of water. The scene appears to be a sunny day with some clouds. I will scan the video for any potential artifacts. My attention is first drawn to the water in the foreground. The surface of the water is not calm; there are visible ripples and reflections from the buildings. However, the texture of the water looks unnatural. Instead of a smooth flow or a natural reflection of light, the surface appears grainy and distorted, with a strange, noisy pattern that doesn't match the movement of the water. This is a clear sign of an artifact. I can identify this as <type>**Texture Jittering**</type> in <t>**[0.0, 5.06]**</t> at <bbox>**[0.0, 0.6, 1.0, 1.0]**</bbox>. Next, I'll examine the buildings in the background. I'm focusing on the cluster of buildings in the center of the frame. As I watch the video, I notice that the structure of the buildings is not stable. The windows and walls seem to flicker and change shape slightly throughout the clip. For example, the green skyscraper on the left appears to warp and the windows on the buildings in the middle seem to shift and distort. This is not a natural change in perspective or lighting. This is a <type>**Structure Anomaly**</type> in <t>**[0.0, 5.06]**</t> at <bbox>**[0.3, 0.0, 0.6, 0.5]**</bbox>. The presence of these two distinct artifacts confirms the video is not authentic.</think> <answer>**Fake**</answer>

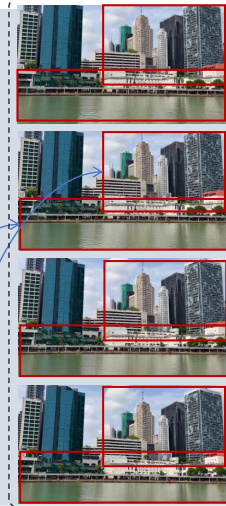


Figure 18. Skyra's Response Example on Fake Videos, Texture Anomaly-Structure Anomaly

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>The video shows a wide shot of a park-like setting with many green trees spaced out across a large, manicured lawn. The camera appears to be static. I am examining the overall visual quality of the scene. My attention is immediately drawn to the color of the grass. It looks unnaturally vibrant and uniform, almost like a bright green carpet rather than natural grass. The color is extremely intense and lacks the subtle variations in shade, texture, and shadow that would be present in a real-world scene. This appears to be a consistent issue throughout the entire clip. This is a clear case of <type>**Color Over-Saturation**</type> in <t>**[0.0, 5.06]**</t> at <bbox>**[0.0, 0.5, 1.0, 1.0]**</bbox>, where the grass is excessively and unnaturally colored. The lack of realistic detail and the overly saturated hue strongly suggest the video is not authentic.</think> <answer>**Fake**</answer>



Figure 19. Skyra's Response Example on Fake Videos, Color & Lighting Anomaly-Color Over-Saturation

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>The video shows a person from a first-person perspective, driving a red convertible car on a suburban street. The person is wearing a red cap with a Ferrari logo. The car is moving forward, and the camera is positioned inside the vehicle, looking out through the front windshield. I am scanning the video for any visual artifacts. The overall scene appears stable, but I'm paying close attention to the movement of objects. As I watch the video progress, I notice something strange happening in the background. While the car is moving forward, the trees and houses on both sides of the road are not static. They seem to be jittering and shaking in a very unnatural, high-frequency manner. This is not consistent with the smooth motion of a car on a road. This erratic movement is a clear sign of an artifact. I can identify this as <type>Camera Motion Inconsistency</type> in <t>[0.49, 4.85]</t> at <bbox>[0.1, 0.1, 0.9, 0.8]</bbox>, where the background elements exhibit a rapid, jittery vibration that does not match the expected motion of the vehicle. This makes the video feel unstable and generated.</think><answer>Fake</answer>

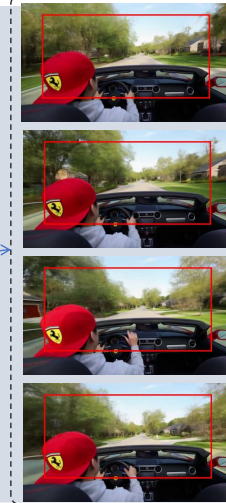


Figure 20. Skyra's Response Example on Fake Videos, Move Forgery-Camera Motion Inconsistency

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>The video shows a chef in a professional kitchen, standing behind a counter with various ingredients and bowls laid out. The chef is wearing a traditional white uniform and a tall white hat. He appears to be giving a presentation or a cooking demonstration. Initially, he is gesturing with his hands while speaking. I'm scanning the video for any visual inconsistencies. The background, lighting, and the chef's face seem generally stable. However, as I continue to watch, my attention is drawn to his hands as he moves them. Around the 2.3-second mark, the chef's hands start to look strange. As he gestures, his hands seem to lose their defined structure. They become blurry and appear to warp, with the fingers and palm blending together in an unnatural way. This is not just motion blur; the actual shape of the hands is deforming. This is a clear example of <type>Shape Distortion</type> in <t>[2.34, 3.41]</t> at <bbox>[0.3, 0.3, 0.8, 0.6]</bbox>. The hands repeatedly lose their proper anatomical form as they move, which is a strong indicator of a generated video.</think><answer>Fake</answer>

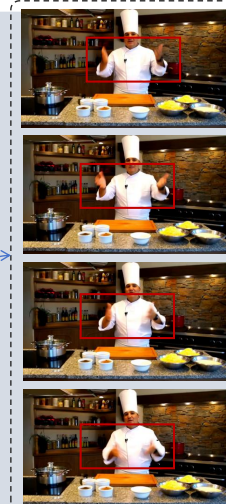


Figure 21. Skyra's Response Example on Fake Videos, Object Inconsistency-Shape Distortion

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>The video shows a first-person perspective of a person drinking from a plastic water bottle. The person is wearing a black jacket and glasses. The setting appears to be indoors, with ceiling lights visible. I am analyzing the video for any signs of digital manipulation. The overall scene seems plausible at first glance. I'll focus on the action of drinking. The person brings the bottle to their mouth and takes several sips. As I watch the sequence, I notice something strange happening around the person's mouth area. Starting around the 1.3-second mark, the person's mouth and the bottle's opening seem to blend together in an unnatural way. The boundary between the person's skin, the lips, and the plastic of the bottle becomes indistinct and wobbly. This is not how two solid objects should interact. This is a clear case of **Abnormal Rigid-Body Crossing** in <t>[1.29, 2.41]</t> at <bbox>[0.4, 0.0, 0.6, 0.3]</bbox>, where the person's mouth and the bottle appear to merge and pass through each other as the person drinks. This physical impossibility confirms the video is not authentic.</think>
<answer>**Fake**</answer>

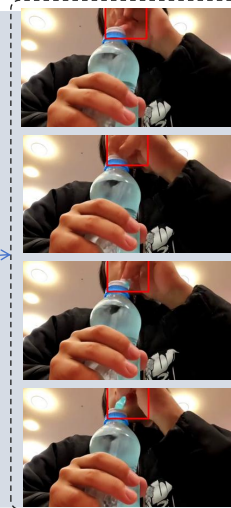


Figure 22. Skyra's Response Example on Fake Videos, Interaction Inconsistency-Abnormal Rigid-Body Crossing

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>The video shows a sports team, dressed in green and white uniforms, performing what appears to be a pre-game ritual or huddle on a large, grassy field in front of a massive crowd. I'm analyzing the movements of the players for any inconsistencies. Initially, the players are mostly stationary, facing each other. As the video progresses, there is a lot of motion. However, the movements seem strange and do not follow the rules of physics or common sense. Players on the left side of the frame, particularly the one in the blue jersey, seem to glide and float into position rather than running or jumping. The overall choreography is complex, but the individual player movements are disjointed and don't seem to have a clear, logical purpose. For example, the player in the center wearing number 10 seems to phase through other players or simply appear in different positions without a natural cause. The entire sequence feels like a poorly animated dance rather than a real athletic performance. This is a clear case of **Unnatural Human Movement** in <t>[0.0, 5.06]</t> at <bbox>[0.0, 0.3, 1.0, 0.8]</bbox>, as the players' running, jumping, and interacting motions are illogical and physically impossible.</think><answer>**Fake**</answer>



Figure 23. Skyra's Response Example on Fake Videos, Unnatural Movement-Unnatural Human Movement

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



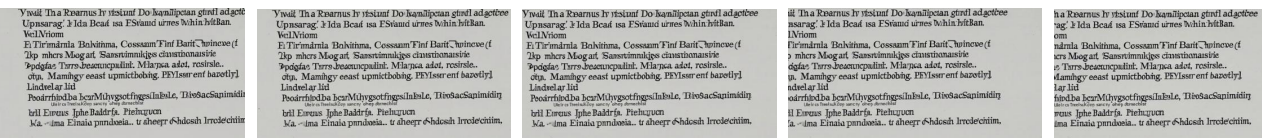
Skyra

<think>The video shows a close-up of a glass mug filled with a yellowish liquid, likely tea, resting on a white wooden surface. A spoon is already in the liquid. I'll watch the sequence of events. The scene is static for the first couple of seconds. Then, starting around the 3-second mark, a stream of liquid is poured into the mug from above. As this new liquid enters the mug, I notice something very strange. The surface of the tea inside the mug ripples and distorts in a way that doesn't seem physically correct. The ripples appear to travel through the liquid and even seem to form and cancel each other out in mid-air, which is impossible. This is a clear **Violation of Physical Law** in **[3.05, 4.35]** at **[0.4, 0.1, 0.7, 0.7]**, as the liquid's behavior defies the physics of fluid dynamics. The way the ripples form and cancel is not how a real liquid would behave when poured into another container.</think>
 <answer>**Fake**</answer>



Figure 24. Skyra's Response Example on Fake Videos, Violation of Causality Law-Violation of Physical Law

System Prompt: You're an ...Your primary task is to review a sequence of video frames and provide a step-by-step analysis of their authenticity.
User Prompt: Here are the video frames... Please analyze the video frames, determine if the video is **real** or **fake**, and provide your reasoning



Skyra

<think>The video displays a static screen filled with lines of text. The characters appear to be a mix of English letters and other symbols, but they form no recognizable words or language. The text is arranged in multiple paragraphs. I am examining the video for any signs of artificial generation. The overall appearance is consistent, with no camera movement. I will now focus on the content of the text. As I scan through the frames, I notice that while the characters within a single line of text seem stable, the lines themselves are not perfectly straight. They exhibit a subtle, high-frequency jitter, as if wobbling or warping slightly. This is particularly noticeable at the beginning of the video. This is a clear sign of an artifact. I have identified **Text Distortion** in **[0.0, 1.27]** at **[0.1, 0.1, 0.9, 0.9]**, where the lines of text are unnaturally bent and unstable. This distortion persists for the first couple of seconds, confirming the artificial nature of the text.</think>
 <answer>**Fake**</answer>

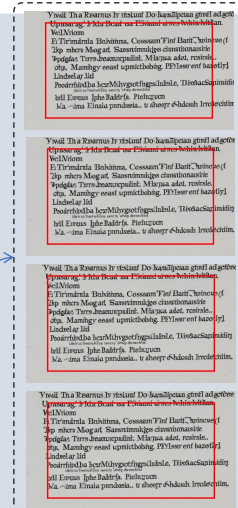


Figure 25. Skyra's Response Example on Fake Videos, Violation of Commonsense-Text Distortion