Thinking with Frames: Generative Video Distortion Evaluation via Frame Reward Model

Supplementary Material

A. Detailed Taxonomy of Structural Distortion

Generative videos typically contain multiple interacting objects, therefore, we construct our taxonomy of structural distortions based on abnormalities in object appearance and object interaction within the video. We categorize structural distortions into two major groups: abnormal object appearance and abnormal object interaction. As illustrated in the section 3.1, the former is further divided according to object characteristics into animal-centric, non-animal-centric, and motion-blur-related distortions. The animal-centric category includes limb deformation, extra limbs, limb incompleteness, torso deformation, and facial deformation. The non-animal-centric category corresponds to non-animal collapse and distortion. Abnormal object interaction primarily refers to mesh penetration. The complete taxonomy is illustrated in Fig. 3. Detailed definitions of each category are provided below:

- **Limb Deformation:** Abnormal distortion of the limbs (arms, hands, legs, feet) of an animal-like subject (including humans, animals, anthropomorphic characters, etc.), violating anatomical plausibility. This may manifest as unnatural bending, merging, or posture misalignment, *e.g.*, hyper-extended or reversed joints, twisted or fused fingers, abnormal stretching of arms, etc. In Fig. 3, the subject's fingers are severely twisted and lose their normal shape and contour, which is representative of limb deformation.
- **Limb Incompleteness:** Partial absence of limbs in the generated subject, such as missing a hand, finger, or leg.
- **Extra Limbs:** The appearance of redundant limbs, *e.g.*, a human with three arms, more than two legs, or more than five fingers. As shown in the second row and first column of Fig. 3, the woman displays anatomically implausible limb duplication, with no proper hands and only an arm remaining on her left side.
- **Torso Deformation:** Abnormal structure or posture of the body's axial region (head, neck, thorax, abdomen, pelvis). Issues include deformation, malformation, absence, redundancy, or unnatural poses, *e.g.*, severely bent waist, head twisted at extreme angles, body discontinuity. In Fig. 3, the woman's head and back are positioned at an impossible angle, which can be categorized as torso deformation.
- **Facial Deformation:** Abnormalities in the face (facial contours and features). Includes facial distortion, missing features, redundant features, or distorted features, *e.g.*, missing mouth, distorted proportions, or multiple overlapping faces. As shown in Fig. 3, the facial deformation refers to a distorted face that lacks normal anatomical structure and contour.
- **Mesh Penetration:** Physical penetration between otherwise independent objects, *e.g.*, an arm intersecting with the torso, a leg passing through a chair, clothing or props penetrating the skin. As an example, two men sitting on a chair in Fig. 3 appear to penetrate through the wire mesh, which is physically impossible.
- **Non-Animal Distortion and Collapse:** Severe distortion, collapse, or unrealistic structural failure affecting non-animal subjects (plants, inanimate objects, or static structures), producing implausible or broken appearances, such as the blurred and collapsed car front shown in the third row and second column of Fig. 3.
- **Motion Blur:** frame blur or trailing artifacts caused by subject motion or generative errors, resulting in unclear boundaries similar to long-exposure camera artifacts.

In addition to the above definitions, we further clarify the anatomical scope used throughout this taxonomy. The face includes both the facial contour and all facial features; limbs include arms, legs, hands, and feet; and the torso encompasses the head, neck, thorax, abdomen, and pelvis. For animals without limbs (*e.g.*, snakes, fish) or stylized characters, all non-facial regions are considered part of the torso. Moreover, we do not treat abnormal posture as a standalone category. Instead, posture-related distortions affecting the axial region are classified as torso deformation, while posture anomalies occurring in the limbs fall under limb deformation.

B. Dataset Annotation Rules

To construct the annotated dataset that forms the foundation of our REACT framework, we collect a large-scale set of frame pairs following the procedure described in Section 3.1 and annotate them according to the taxonomy detailed in Appendix A. The annotation process comprises three components: (1) distortion recognition, (2) spatial grounding of each distortion label for every frame, and (3) human preference annotation, which we denote as GSB (*i.e.*, Good / Same / Bad). Specifically, given a frame pair, annotators first examine each frame individually and assign bounding boxes corresponding to all annotated distortion types (*i.e.*, attribution labels). They then determine a preference judgment for the pair based on the number and severity of the annotated bounding boxes and their associated attribution labels. To ensure consistency and reliability in eval-

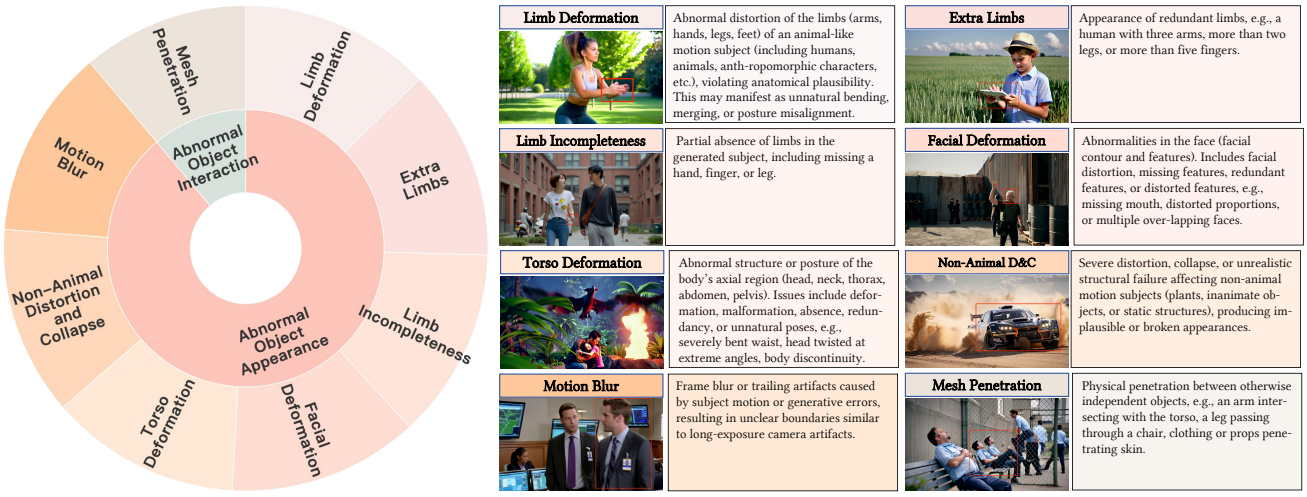

Figure 3. **Detailed Explanation of Our Proposed Taxonomy of Structural Distortions in Generative Videos.** Representative examples for each distortion category are also provided.

uating structural distortions in generative videos, we establish detailed annotation guidelines for all three components.

For the distortion recognition task, annotators may assign at most three issue labels from the taxonomy to each frame. When a frame exhibits more than three issues, the selection is based primarily on the spatial extent and perceptual severity of the defects. For the grounding task, multiple bounding boxes may be assigned to a single attribution label when the corresponding distortion appears in multiple disjoint regions. Each bounding box must fully encompass the relevant distorted region such that the problematic content can be identified solely from information within the box, without relying on external context. When occlusion occurs, annotators approximate the full spatial extent of the affected area. In conclusion, bounding boxes should avoid unnecessary inclusion of irrelevant visual content to minimize interference from unrelated structures. For the human preference task, the frame containing fewer attribution labels and bounding boxes is preferred. A Same preference is assigned only when (1) both frames exhibit the same distortion types with comparable severity, or (2) neither frame contains identifiable structural distortion issues. Certain special cases follow additional principles outlined below:

- **Prioritizing Animal-Centric Labels.** When more than three structural distortion types occur in a frame, animal-centric labels, textitlimb deformation, extra limbs, limb incompleteness, torso deformation, and facial deformation, are prioritized. Non-animal collapse and distortion and mesh penetration follow, while *motion blur* is considered last. This prioritization also applies to human preference annotation, where animal-centric distortions are treated as more severe in the GSB decision process.
- **Distinguishing Motion Blur from Deformation and Collapse.** Motion blur or trailing is annotated only when the subject displays explicit motion cues and retains an otherwise coherent and correct outline, with blurring localized around the moving edges. Blur, tearing, or deformation occurring in static objects (*e.g.*, buildings, vegetation, background regions), *i.e.*, non-animal entities under our taxonomy, is consistently attributed to non-animal collapse and distortion.
- **Distinguishing Limb Incompleteness from Limb Deformation.** Limb incompleteness is assigned when a limb component is entirely or partially absent, such as missing hands or feet, fewer than five fingers, or fully missing limbs. When a limb is present but structurally collapsed due to distortion, the appropriate label is limb deformation rather than limb incompleteness.

C. Prompt Templates

In this section, we provide a clear overview of the prompts used throughout the entire process. We first introduce the prompt designed for efficient CoT synthesis, as shown in Fig. 7. Specifically, we supply the annotated attribution labels together with their corresponding bounding boxes, and instruct Gemini to simulate the reasoning process that leads to these labels and bounding boxes. For structural distortion evaluation, we design two types of prompts based on our proposed taxonomy: one for the human preference alignment task and the other for the distortion recognition task. The prompt for human preference alignment is shown in Fig. 4, while the prompt for distortion recognition is presented in Fig. 5. By incorporating detailed explanations of each distortion category, these prompts enable REACT to develop a more comprehensive understanding of structural distortions in generative videos, thereby producing more ac-

curate evaluation results.

D. Additional Experiments Results

D.1. Evaluation Prompt

When evaluating human preference alignment with REACT-Video, we apply each video reward model, VideoScore2, UnifiedReward, and VideoReward, using their original prompts, which are designed to assess multiple aspects of video quality holistically. For general MLLMs, we adopt the same prompt used in REACT, which includes detailed descriptions of each distortion type and the principles for assigning point-wise scores. This prompt guides the models to generate distortion-aware point-wise quality assessments.For image evaluators, we use their native prompts and further introduce the REACT prompt as a refined supplementary prompt, allowing these models to incorporate auxiliary knowledge about structural distortions in generative videos during the additional experiments.

When evaluating distortion recognition with REACT-Frame, only image evaluators and general MLLMs are responsible for this task. All models, including MagicAccessor, are instructed using the prompt shown in Fig. 5, which contains detailed explanations of all attribution labels associated with structural distortion. This is because all these models are trained or adapted from general-purpose MLLMs capable of instruction following, enabling them to perform the required annotation tasks under a well-specified prompt.

D.2. Evaluation Metrics

For the human preference alignment evaluation, we use preference accuracy as the metric to assess the performance of REACT. Specifically, we report accuracy with tie and without tie. Accuracy without tie directly compares the point-wise scores of the two frames in each pair and assigns the preference to the frame with the higher score. For accuracy with tie, we additionally consider the cases where the two frames are essentially equivalent, that is, if the score difference between the two frames falls below a predefined threshold, the pair is treated as a tie. Since all baselines are prompted to produce point-wise scores rather than explicitly comparing the frame pairs, we first convert their point-wise scores into pairwise preferences following the above procedure. As described in Section 4.2, we compute the VQ score and MQ score, and their combined overall score, to derive the final preference for video evaluators. For VideoReward, VQ and MQ correspond to the "visual quality" and "motion quality" dimensions, respectively. For VideoScore, VQ corresponds to "visual quality" and MQ corresponds to "physical/common-sense consistency". For UnifiedReward, VQ maps to "visual quality," while MQ is defined as the average of "temporal consistency" and "factual consistency".

Table 5. **Additional Experiments on GenAI Benchmark and VideoGen-RewardBench.**

Model	VideoGen-RewardBench				GenAI	
	VQ		MQ			
	Acc w/ Tie	Acc w/o Tie	Acc w/ Tie	Acc w/o Tie	Acc w/ Tie	Acc w/o Tie
VideoScore2	0.424	0.515	0.383	0.706	0.391	0.616
UnifiedReward	<u>0.589</u>	<u>0.701</u>	<u>0.475</u>	<u>0.749</u>	**0.548**	<u>0.709</u>
VideoReward	**0.660**	**0.746**	**0.596**	**0.756**	<u>0.491</u>	**0.728**
Qinsight	0.367	0.533	0.372	0.663	0.376	0.571
Our REACT	0.402	0.538	0.386	0.626	0.376	0.581

For the distortion recognition task, we evaluate the performance of REACT using precision, recall, and F1-score, which measure how accurately the model identifies frames suffering from structural distortions. The calculation is defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{10}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{11}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. Precision reflects the accuracy of positive predictions, i.e., the proportion of predicted positive samples that are truly positive. Recall reflects the coverage of the model, i.e., the proportion of true positive samples that are correctly identified. F1-score provides a comprehensive measure of overall performance by balancing precision and recall.

D.3. Additional Human Preference Alignment

We also conduct experiments on the GenAI benchmark and VideoGen-RewardBench. The former is a reward benchmark for generative models, annotated with human preferences over visual content produced by image editing, image generation, and video generation models. We use the subset corresponding to generative video to evaluate the performance of our REACT on video quality assessment. The latter benchmark extends VideoGen-Eval to construct a human-preference dataset for evaluating reward models on modern text-to-video (T2V) models. As shown in Table 5, REACT is slightly inferior to video-based evaluators in terms of overall preference accuracy. We attribute this to the fact that REACT is grounded in a new preference formulation that emphasizes structural distortions—an aspect not explicitly modeled in existing video evaluation methods. Nevertheless, REACT outperforms the image-based evaluator Q-Insight, demonstrating its stronger ability to assess generative video quality.

Text Prompt for Our REACT in Human Preference Alignment Task

What is your overall rating on the visual quality of this frame? The rating should be a floating-point number between 1 and 5, rounded to two decimal places. A rating of 1 represents very poor visual quality, and a rating of 5 represents excellent visual quality. The visual quality issues to be considered include the following:

- **Limb Deformation:** Abnormal distortion of the limbs (arms, hands, legs, feet) of an animal-like subject (including humans, animals, anthropomorphic characters, etc.), violating anatomical plausibility. This may manifest as unnatural bending, merging, or posture misalignment, e.g., hyper-extended or reversed joints, twisted or fused fingers, abnormal stretching of arms, etc.
- **Limb Incompleteness:** Partial absence of limbs in the generated subject, such as missing a hand, finger, or leg.
- **Extra Limbs:** The appearance of redundant limbs, e.g., a human with three arms, more than two legs, or more than five fingers.
- **Torso Deformation:** Abnormal structure or posture of the body's axial region (head, neck, thorax, abdomen, pelvis). Issues include deformation, malformation, absence, redundancy, or unnatural poses, e.g., severely bent waist, head twisted at extreme angles, body discontinuity.
- **Facial Deformation:** Abnormalities in the face (facial contours and features). Includes facial distortion, missing features, redundant features, or distorted features, e.g., missing mouth, distorted proportions, or multiple overlapping faces.
- **Mesh Penetration:** Physical penetration between otherwise independent objects, e.g., an arm intersecting with the torso, a leg passing through a chair, clothing or props penetrating the skin.
- **Non-Animal Distortion and Collapse:** Severe distortion, collapse, or unrealistic structural failure affecting non-animal subjects (plants, inanimate objects, or static structures), producing implausible or broken appearances.
- **Motion Blur:** frame blur or trailing artifacts caused by subject motion or generative errors, resulting in unclear boundaries similar to long-exposure camera artifacts.

Please first assess whether the frame exhibits any of the issues listed above, and then provide an overall rating for the picture. The final answer should be returned in JSON format with the following keys:

```
{
    "Attribution labels": [A list of the detected issues or "null" if none are found],
    "rating": [The score]
}
```

Figure 4. **Text Prompt for Our REACT in Human Preference Alignment Task.**

D.4. Performance on Improving Video Generation

To further demonstrate the effectiveness of REACT in improving the visual quality of generated videos, we integrate it into two representative paradigms, Best-of-N sampling and Flow-DPO [27], on the open-source video generation model Wan-2.1-1.3B [43], and compare it against state-of-the-art reward models on VBench[11]. For Best-of-N sampling, we generate five videos for each prompt and select the one with the highest reward score. For Flow-DPO, we sample 5.7K prompts from the training dataset and generate videos with Wan-2.1-1.3B, where the positive and negative samples are determined according to the reward scores assigned by the corresponding reward model.

As shown in Tab. 6, under Best-of-N sampling, REACT alone achieves performance competitive with UnifiedReward, slightly outperforming it in Imaging Quality and Aesthetic Quality while maintaining comparable results on Background Consistency and Subject Consistency. These results indicate that REACT can effectively improve the visual fidelity of generated videos. Under Flow-DPO post-training, REACT further surpasses UnifiedReward in Imaging Quality and Aesthetic Quality, demonstrating that accurate assessment of structural distortions provides a more reliable supervision signal for video generation

Furthermore, we evaluate a simple reward fusion strategy that combines REACT and UnifiedReward by averaging their scores as the final reward for generated videos. This combined model yields additional gains in both paradigms and achieves the best performance across all evaluated metrics. These results suggest that REACT captures structural cues that are complementary to existing reward models, and that incorporating such feedback can further improve overall video generation quality.

D.5. Case Study

We present qualitative results in Fig. 6. In the first row, the video contains severe structural distortions, and our REACT successfully identifies all distortions and assigns a reliable point-wise score reflective of its low visual quality. In contrast, the second row shows a high-quality video without structural distortions. Likewise, REACT correctly recognizes it as a normal video and provides a correspondingly

Text Prompt for Our REACT in Distortion Recognition Task

Analyze the provided frame to determine whether it exhibits any of the following visual quality issues: Limb Deformation, Torso Deformation, Facial Deformation, Limb Incompleteness, Extra Limbs, Mesh Penetration, Non-animal Distortion and Collapse, Motion Blur.

- **Limb Deformation:** Abnormal distortion of the limbs (arms, hands, legs, feet) of an animal-like subject (including humans, animals, anthropomorphic characters, etc.), violating anatomical plausibility. This may manifest as unnatural bending, merging, or posture misalignment, e.g., hyper-extended or reversed joints, twisted or fused fingers, abnormal stretching of arms, etc.
- **Limb Incompleteness:** Partial absence of limbs in the generated subject, such as missing a hand, finger, or leg.
- **Extra Limbs:** The appearance of redundant limbs, e.g., a human with three arms, more than two legs, or more than five fingers.
- **Torso Deformation:** Abnormal structure or posture of the body's axial region (head, neck, thorax, abdomen, pelvis). Issues include deformation, malformation, absence, redundancy, or unnatural poses, e.g., severely bent waist, head twisted at extreme angles, body discontinuity.
- **Facial Deformation:** Abnormalities in the face (facial contours and features). Includes facial distortion, missing features, redundant features, or distorted features, e.g., missing mouth, distorted proportions, or multiple overlapping faces.
- **Mesh Penetration:** Physical penetration between otherwise independent objects, e.g., an arm intersecting with the torso, a leg passing through a chair, clothing or props penetrating the skin.
- **Non-Animal Distortion and Collapse:** Severe distortion, collapse, or unrealistic structural failure affecting non-animal subjects (plants, inanimate objects, or static structures), producing implausible or broken appearances.
- **Motion Blur:** frame blur or trailing artifacts caused by subject motion or generative errors, resulting in unclear boundaries similar to long-exposure camera artifacts.

If any issues are detected, identify the three most severe ones. Return the result in JSON format with the following keys:

```
{
    "Attribution labels": [A list of the detected issues or "null" if none are found]
}
```

Figure 5. **Text Prompt for Our REACT in Distortion Recognition Task.**

high score. These qualitative examples clearly demonstrate that REACT performs well in distortion evaluation, both in accurately recognizing structural distortions and in assigning reliable point-wise scores.

Table 6. **Comparison of Reward Models for Improving Video Generation Quality on VBench.** Our REACT substantially improves video generation quality, and integrating it with other SOTA reward models yields additional gains.

Model	VBench				
	Background Consistency ↑	Dynamic Degree ↑	Imaging Quality ↑	Subject Consistency ↑	Aesthetic Quality ↑
Wan-2.1-1.3B	0.951	0.527	0.649	0.948	0.522
w/ Best-of-N					
UnifiedReward (UR)	0.957	0.541	0.674	0.959	0.542
REACT	0.955	0.527	0.675	0.955	0.547
UR+REACT	0.957	0.541	0.675	0.960	0.547
w/ Flow-DPO					
UnifiedReward (UR)	0.971	0.542	0.690	0.977	0.547
REACT	0.963	0.536	0.691	0.977	0.549
UR+REACT	**0.981**	**0.554**	**0.694**	**0.998**	**0.550**

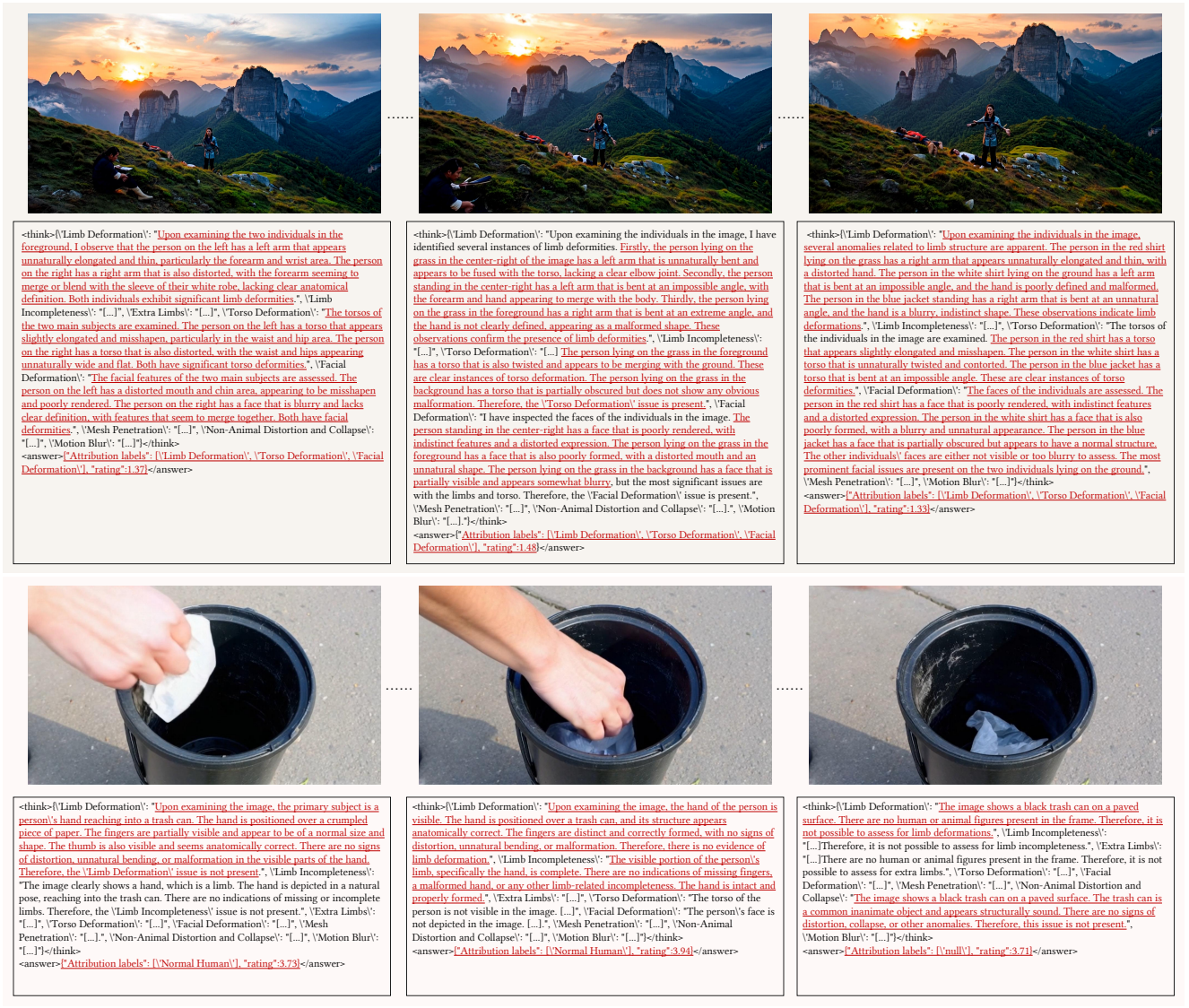

Figure 6. **Case Study of REACT for Distortion Evaluation in Generative Videos.** The two presented video cases illustrate that REACT effectively identifies structural distortions and produces reliable point-wise assessments for generative videos.

Text Prompt for CoT Synthesis

Role and Goal

You are an expert in generated frame quality assessment.

You are given an frame that may have **dynamic quality issues**, **along with a set of annotations `"<label with bbox>"` (each item pairs an attribution `"<label>"` with a corresponding bounding box `"<bbox>"`)**.

Annotation definitions:

- **`<label>`:** List[choice]
  Each entry denotes a dynamic quality issue present in the frame. Candidate labels include: ***limb deformation, limb incompleteness, extra limbs, torso deformation, facial deformation, mesh penetration, non-animal distortion and collapse, motion blur, and no issue.***
- **`<bbox>`:** List[list]
  Each entry is $[x_1, y_1, x_2, y_2]$, where:(1) x_1**:** x-coordinate of the top-left corner of the bounding box;(2) y_1**:** y-coordinate of the top-left corner of the bounding box;(3) x_2**:** x-coordinate of the bottom-right corner of the bounding box;(4) y_2**:** y-coordinate of the bottom-right corner of the bounding box.
- **`<label with bbox>`:** List[tuple]
  Each item consists of an attribution label $<$ label $>$ and its corresponding bounding box $<$ bbox $>$.

Task description:

Your task is: **Assume you don't know the content of these labels. Based only on visual features you observe in the frame, analyze step by step what problems are present**, and ultimately infer the phenomenon corresponding to the attribution label. Bounding box information serves only as a localization reference to help you confirm the problematic area, but it must not drive your judgment. Consider the frame holistically, proceed step by step, and naturally infer the likely attribution label. This process must reflect a professional **Chain of Thought**.

Output requirement:

The final result must be returned as a **complete JSON file**. **Do not output any content or explanatory text outside the JSON.**

Core Instructions

1. **Chain of Thought (CoT)**
   Generate the analysis process corresponding to each label based on the frame itself, meeting the following requirements:
   - Show a typical, professional analysis workflow for generated frame quality assessment, determining whether the frame exhibits any of the following issues: *limb deformation, limb incompleteness, extra limbs, torso deformation, facial deformation, mesh penetration, non-animal distortion and collapse, motion blur, and no issue. Details:*
     - **`Limb Deformation`**: Abnormal distortion of the limbs (arms, hands, legs, feet) of an animal-like motion subject (including humans, animals, anthropomorphic characters, etc.), violating anatomical plausibility. This may manifest as unnatural bending, merging, or posture misalignment, e.g., hyper-extended or reversed joints, twisted or fused fingers, abnormal stretching of arms, etc.
     - **`Limb Incompleteness`**: Partial absence of limbs in the generated subject, e.g., missing a hand, finger, or leg.
     - **`Extra Limbs`**: Appearance of redundant limbs, e.g., a human with three arms, more than two legs, or more than five fingers.
     - **`Torso Deformation`**: Abnormal structure or posture of the body's axial region (head, neck, thorax, abdomen, pelvis). Issues include deformation, malformation, absence, redundancy, or unnatural poses, e.g., severely bent waist, head twisted at extreme angles, body discontinuity.
     - **`Facial Deformation`**: Abnormalities in the face (facial contour and features). Includes facial distortion, missing features, redundant features, or distorted features, e.g., missing mouth, distorted proportions, or multiple overlapping faces.
     - **`Mesh Penetration`**: Physical penetration between otherwise independent objects, e.g., an arm intersecting with the torso, a leg passing through a chair, clothing or props penetrating skin.
     - **`Non-animal Distortion and Collapse`**: Severe distortion, collapse, or unrealistic structural failure affecting non-animal motion subjects (plants, inanimate objects, or static structures), producing implausible or broken appearances.
     - **`Motion Blur`**: frame blur or trailing artifacts caused by subject motion or generative errors, resulting in unclear boundaries similar to long-exposure camera artifacts.
     - **`No Issue`**: The frame has no apparent dynamic quality defects overall.
   - **Source of evidence**: Base reasoning and judgments **only on observable visual features**.
   - **Independence constraint**: **Do not use the attribution labels or their bounding boxes for reverse validation or inference**; they may be used only for comparison after your reasoning is complete.
   - **Factuality**: **Do not fabricate elements that are not present in the frame** (e.g., inventing objects/people/actions).

- You may naturally arrive at the attribution indicated by the labels, but the process must be based on observation rather than hints from labels.
- **Analyze each attribution label one by one, with an independent chain of thought for each label.**

Consistency requirement: The final inferred attribution must match the ground-truth "<label>", and the problematic regions indicated during reasoning must strictly align with "<label with bbox>".

2. **JSON Output Format**
   Your output must be a **clear, syntactically correct, valid JSON object** where each attribution label is a **key**, and the corresponding analysis process is the **value**. **Do not output anything outside the JSON structure. The return must be valid JSON; Markdown styling or pseudo-JSON is strictly forbidden.**
   JSON format:

```
{
  "COT": {
    "Limb Deformation": "The reasoning process determining whether this issue exists in the frame",
    "Limb Incompleteness": "The reasoning process determining whether this issue exists in the frame",
    "Extra Limbs": "The reasoning process determining whether this issue exists in the frame",
    "Torso Deformation": "The reasoning process determining whether this issue exists in the frame",
    "Facial Deformation": "The reasoning process determining whether this issue exists in the frame",
    "mesh penetration": "The reasoning process determining whether this issue exists in the frame",
    "non-animal distortion and collapse": "The reasoning process determining whether this issue exists
        in the frame",
    "Motion Blur": "The reasoning process determining whether this issue exists in the frame"
  },
  "Attribution Label": "Based on the CoT, the label corresponding to the issue that truly exists in the
      frame",
  "Problem Region": "Based on the CoT, the region corresponding to the issue that truly exists in the
      frame"
}
```

Field descriptions

- **COT**: **[To fill]** For the given frame, **analyze and verify each issue label (limb deformation, limb incompleteness, extra limbs, torso deformation, facial deformation, mesh penetration, non-animal distortion and collapse, motion blur) in turn**, determining whether the issue exists and writing out the complete reasoning process for each label in order. If none of these issues appear, you may provide the **"no Issue"** attribution label. Suggested content includes:
  - **Input evidence** (data source, frame/region/timestamp, visible features);
  - **Reasoning steps** (logical transition from evidence to decision and exclusion tests; note: the visual evidence used in reasoning should **fall within the region indicated by "<label with bbox>"**);
  - **Conclusion** (final judgment).
- **Attribution Label**: **[To fill]** The anomaly category inferred from the CoT analysis, **which must strictly match the ground-truth "<label>"**.
- **Problem Region**: **[To fill]** The frame region corresponding to the inferred attribution. **Each anomalous region must match the meaning of the attribution label, and the overall region must strictly match the ground-truth bounding box "<label with bbox>"**.

Please begin your analysis.

Figure 7. **Text prompt for Efficient CoT Synthesis.**