# Beyond Static Artifacts: A Forensic Benchmark for Video Deepfake Reasoning in Vision Language Models

## Supplementary Material

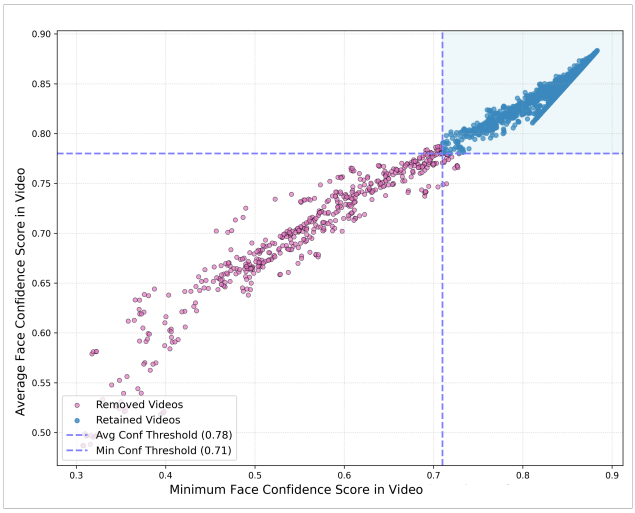


Figure 1. Distribution of Videos across Face Detection Confidence Metrics and the Filtering Boundary. Each point represents a video from the original set. The filtering boundary (dashed light blue lines) separates the retained high-quality samples (blue points) from the removed low-quality samples (light purple points), demonstrating the rigorous exclusion of videos with unstable or weak face detections.

## 1. Data Filtering

We performed frame-level face detection on all video files in the FF++ dataset, and utilized the confidence scores from face detection as the benchmark for data cleaning. A video was deemed satisfactory and thus retained only if its average face confidence score exceeded 0.78 and its minimum per-frame confidence score was above 0.71. This two-pronged criterion ensures not only a high overall quality but, crucially, prevents the inclusion of videos with transient but severe detection failures (e.g., due to heavy occlusion or poor camera angle). As visualized in Figure 1, this process effectively removes approximately 10% of the initial unsatisfactory videos, yielding a refined set of over 4500 high-quality samples.

To empirically validate the efficacy of our data curation strategy, we conducted a sensitivity analysis by varying the retention ratio (100%, 90%, 80%, and 70%) based on the confidence scores of the filtering mechanism. The results are shown in Table 1.

Comparing the 90% setting against the full dataset (100%), we observe a measurable degradation in the latter, particularly in in-domain detection accuracy (e.g., a decline from 73.8% to 71.5% for Qwen2.5-VL). This suggests that the bottom 10% of samples introduce detrimental noise rather than informative features. These low-confidence samples likely contain severe visual degradation or alignment errors, which impair the optimization process by confusing the model's decision boundaries.

Conversely, increasing the filtering strictness to 80% and 70% negatively impacts cross-domain generalization (CDF, DFo, WDF). This decline indicates that while the bottom 10% constitutes noise, the subsequent data segments contribute significantly to the distributional diversity of the training set. Excluding these samples reduces the feature variance necessary for the model to learn robust representations against unseen manipulation types. Consequently, the

Table 1. Ablation study on data filtering ratios. The table reports **(Top)** Overall benchmark accuracy including real cases; and **(Bottom)** In-domain and Cross-domain tasks. **90%** corresponds to our proposed setting.

| *Overall Benchmark Accuracy (%)* | | | | | |
|---|---|---|---|---|---|
| **Model** | **Data Ratio** | **Level 1** | **Level 2** | **Level 3** | **Average** |
| Qwen2.5-VL | 100% (Raw) | **90.3** | 40.1 | 23.5 | 51.3 |
| | 90% (Ours) | 89.9 | **41.4** | **25.8** | **52.4** |
| | 80% | 89.7 | 41.0 | 24.9 | 51.9 |
| | 70% | 88.1 | 39.5 | 23.2 | 50.3 |
| | 60% | 85.4 | 36.8 | 20.1 | 47.4 |
| | 50% | 81.2 | 33.2 | 18.5 | 44.3 |
| LLaVA-NeXT | 100% (Raw) | **89.2** | 43.9 | 24.2 | 52.4 |
| | 90% (Ours) | 88.8 | 45.8 | **26.5** | **53.7** |
| | 80% | 88.5 | **45.9** | 26.1 | 53.5 |
| | 70% | 86.9 | 43.5 | 24.4 | 51.6 |
| | 60% | 83.5 | 40.1 | 21.8 | 48.5 |
| | 50% | 79.4 | 36.5 | 19.6 | 45.2 |

| *In-domain (Avg) & Cross-domain Generalization (Acc)* | | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **Data Ratio** | **In-MCQ** | **In-Det** | **CDF** | **DFo** | **WDF** |
| Qwen2.5-VL | 100% (Raw) | 40.2 | 71.5 | 71.5 | 76.2 | 63.8 |
| | 90% (Ours) | **41.5** | 73.8 | **73.3** | **78.6** | **65.9** |
| | 80% | 41.7 | **74.2** | 72.8 | 78.1 | 64.9 |
| | 70% | 39.8 | 72.1 | 69.5 | 76.4 | 62.5 |
| | 60% | 36.5 | 67.2 | 66.1 | 72.8 | 58.9 |
| | 50% | 33.2 | 60.8 | 62.4 | 68.5 | 55.1 |
| LLaVA-NeXT | 100% (Raw) | 43.1 | 67.5 | 70.8 | 75.9 | 61.2 |
| | 90% (Ours) | **44.4** | **69.3** | 72.9 | **77.8** | **63.1** |
| | 80% | 44.2 | 69.0 | **73.1** | 77.2 | 62.5 |
| | 70% | 42.5 | 66.8 | 69.1 | 74.8 | 60.4 |
| | 60% | 39.1 | 62.4 | 65.3 | 70.5 | 57.2 |
| | 50% | 35.8 | 58.1 | 61.7 | 66.2 | 53.8 |

90% threshold establishes an optimal equilibrium between maximizing data quality and preserving the semantic diversity required for effective generalization.

## 2. Prompt Details

To facilitate reproducibility and provide comprehensive implementation details, we present the exact prompts employed to guide the LLM in our pipeline.

**Prompt for Description Parsing.** This prompt guides the LLM to decompose raw video descriptions into atomic, structured annotations.

Listing 1. Prompt for Description Parsing

```
You are a data processing assistant.
Your task is to extract fine-grained "
    atomic annotations" from a raw video
    description.

Input Description:
"{raw_description}" (e.g., "The video shows
     a person whose nose looks blurry and
    the mouth area has inconsistent skin
    textures.")

Instructions:
1. Decompose the description into separate
    atomic units.
2. Each unit must describe exactly ONE
    facial region and ONE specific artifact
    type.
3. Map the facial region to one of these
    categories: [Eyes, Nose, Mouth, Jaw,
    Ears].
4. Map the artifact type to standard
    forensic terms: [Blur, Color Distortion,
     Texture Inconsistency, Boundary
    Artifacts, Lighting Anomaly].
5. Ignore subjective adjectives (e.g., "
    weird", "ugly") and focus on visual
    evidence.

Output Format (JSON):
[
  {"region": "Nose", "artifact": "Blur"},
  {"region": "Mouth", "artifact": "Texture
      Inconsistency"}
]
```

**Prompt for Generating QA in Level 1.** This prompt constructs Region Perception and Edge Perception tasks with visually plausible distractors.

Listing 2. Prompt for Generating QA in Level 1

```
You are a Visual QA generator. Generate a
    binary choice question (A/B) to test the
     perception of artifacts in a specific
    facial region.

Input Data:
- Region: "{region}" (e.g., Skin)
- State: "{state}" (e.g., Pixelated / Real)

Instructions:
1. Question: Must start with "Observe the {
    region}." followed by "Which option best
     describes its visual quality?"
2. Options:
   - Option A: Describe a high-quality,
       natural appearance (e.g., "Clear,
       well-defined...").
   - Option B: Describe the artifact or
       distortion (e.g., "Pixelated, blurry,
        or fake-looking").
   - Randomize the order of A and B in the
       final output.

Example Output:
{
  "question": "Observe the Skin. Which
      option best describes its visual
      quality?",
  "options": {
    "A": "Clear, well-defined, with natural
         appearance.",
    "B": "Blurred, distorted, or fake-
        looking."
  },
  "answer": "B"
}
```

**Prompt for Generating QA in Level 2.** This prompt generates Type Understanding, Region Grounding, and Temporal Grounding tasks by masking one dimension of the annotation.

Listing 3. Prompt for Generating QA in Level 2

```
You are a generic QA generator. Given a
    video timeline and artifact details,
    generate a temporal grounding question.

Input Data:
- Artifact: "{artifact}"
- Manipulation Time: "{start}s-{end}s" (or
    "None" if Real)
- Task: "Identify the time range of the
    manipulation."

Instructions:
1. Question: "During which time range does
    the face appear most realistic, without
    manipulation artifacts?" (or inverse for
     fake).
2. Options (Generate 5 options A-E):
   - Include the Correct Time Range.
   - Include 3 Distractor Time Ranges (non-
```

```
      overlapping or partial).
   - ALWAYS include one "Authentic" option:
       "The video appears harmonious and
      consistent, showing no signs of
      manipulation. It is authentic."

Example Output:
{
  "question": "During which time range does
      the face appear most realistic,
     without manipulation artifacts?",
  "options": {
    "A": "8.40s-10.72s",
    "B": "10.72s-13.44s",
    "C": "3.80s-7.04s",
    "D": "The video appears harmonious and
       consistent, showing no signs of
       manipulation. It is authentic.",
    "E": "10.44s-12.88s"
  },
  "answer": "D"
}
```

**Prompt for Generating QA in Level 3.** This prompt constructs complex options combining temporal, spatial, and artifact attributes.

Listing 4. Prompt for Generating QA in Level 3

```
You are a Forensic Expert. Generate a
    multiple-choice question where the
    options are detailed forensic
    descriptions.

Input Trace:
- Correct: Time="{t_start}-{t_end}", Region
    ="{region}", Artifact="{type}"

Instructions:
1. Question: "Which of the following
    descriptions best matches the
    manipulated content of this video
    segment?"
2. Option Format: MUST follow this template
    : "Between {t_start} and {t_end}, the {
    region} shows signs of {type}."
3. Distractors:
   - Generate options that mix up the
      dimensions (e.g., Correct Time but
      Wrong Region, or Correct Region but
      Wrong Type).
   - Include an "Authentic" option if
      applicable.

Example Output:
{
  "question": "<video>Which of the
     following descriptions best matches
     the manipulated content of this video
      segment?",
  "options": {
    "A": "Between 4.44s and 4.96s, the
       Mouth shows signs of Blurred edges
       .",
    "B": "The visual sequence is continuous
        and believable... It looks real.",
    "C": "Between 5.29s and 5.87s, the Ears
        shows signs of Pixelation.",
    "D": "Between 4.44s and 4.96s, the Ears
        shows signs of Pixelation."
  },
  "answer": "D"
}
```

## 3. Modeling Details

The models and corresponding parameters used for training and evaluation are presented in Table 2.

Table 2. Hyperparameters for Training and Inference

| Configuration | Value |
|---|---|
| *Video Input Preprocessing* | |
| Max Pixels per Video | 50176 |
| Max Frames per Video | 36 |
| Input Video Frame Resolution | $448 \times 448$ |
| *Training / Fine-tuning Configuration* | |
| Total Training Epochs | 1 |
| Optimizer | AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$) |
| Weight Decay | 0.1 |
| Global Batch Size | 16 |
| Batch Size (per GPU) | 4 |
| Total GPUs Used | 4 |
| Learning Rate (LR) | $1 \times 10^{-5}$ |
| LR Scheduler | Cosine Decay |
| Warm-up Steps | 500 |

## 4. Annotation Platform

We have developed an online human annotation platform and recruited volunteers within the organization, all holding at least a bachelor's degree, to perform the annotations. The annotation platform is displayed as shown in Figure 2.

## 5. Verification Platform

Following the annotation process section 4, we have established an online data validation platform to assess the usability of the QA data we have created. The verification platform is displayed as shown in Figure 3, Figure 4 and Figure 5. For each level, we have designed distinct verification processes.

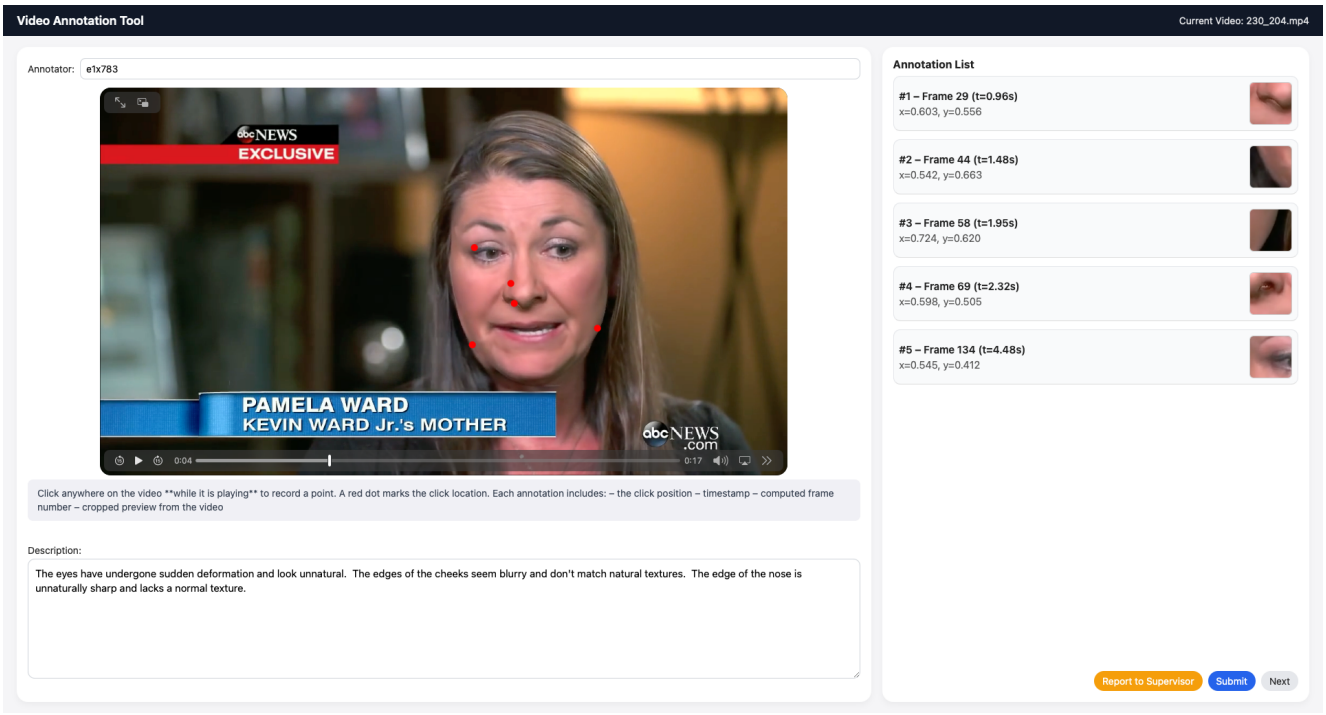


Figure 2. Overview of our custom-developed video annotation interface. The platform is designed to capture fine-grained spatiotemporal forgery traces. The central player allows annotators to mark dynamic artifacts (indicated by red dots) in real-time, recording precise $(x, y, t)$ coordinates. The right panel displays the sequential list of annotated clicks with corresponding frame previews, while the bottom text field captures a detailed natural language description of the observed manipulation anomalies.

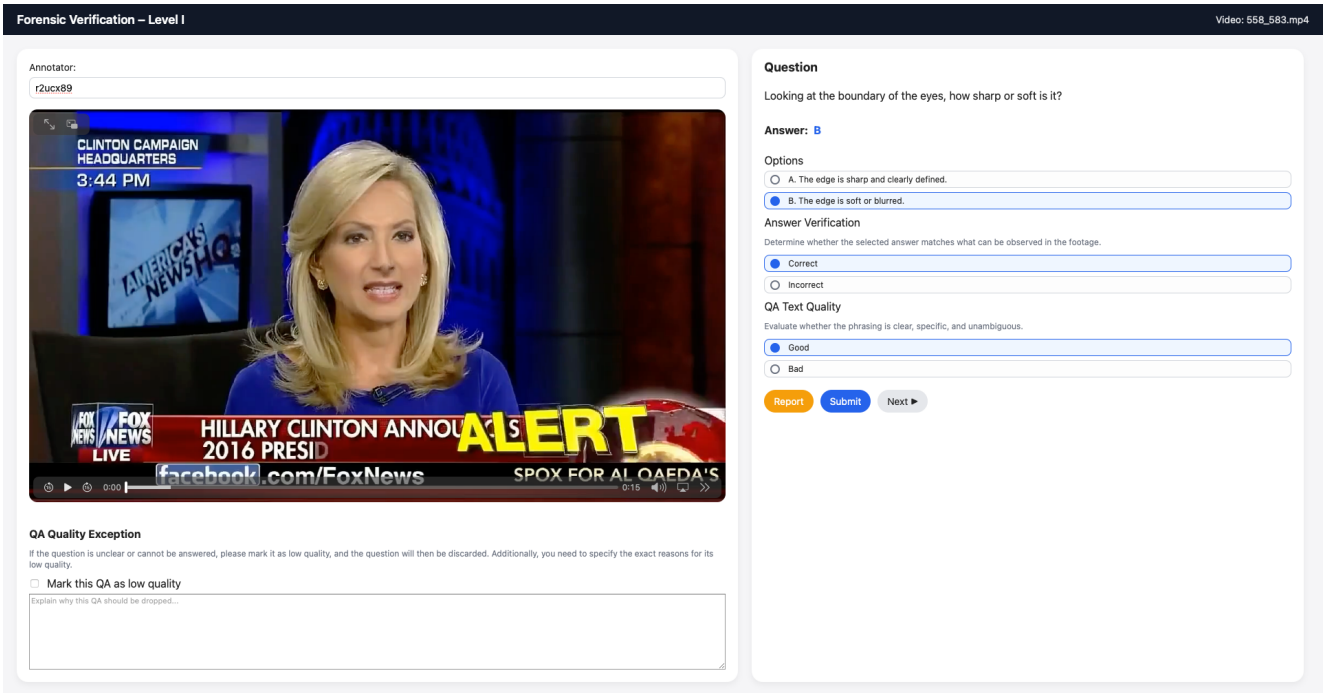


Figure 3. User interface of our verification platform for Level 1 (Facial Perception). To ensure high data quality, validators execute a dual-check protocol: (1) Answer Verification, ensuring the selected option (e.g., "soft or blurred edges") accurately reflects the visual artifacts observed in the video; and (2) Text Quality Assessment, confirming that the phrasing is unambiguous. The "Exception" panel at the bottom allows validators to flag and discard samples with severe hallucinations or low-quality generation.

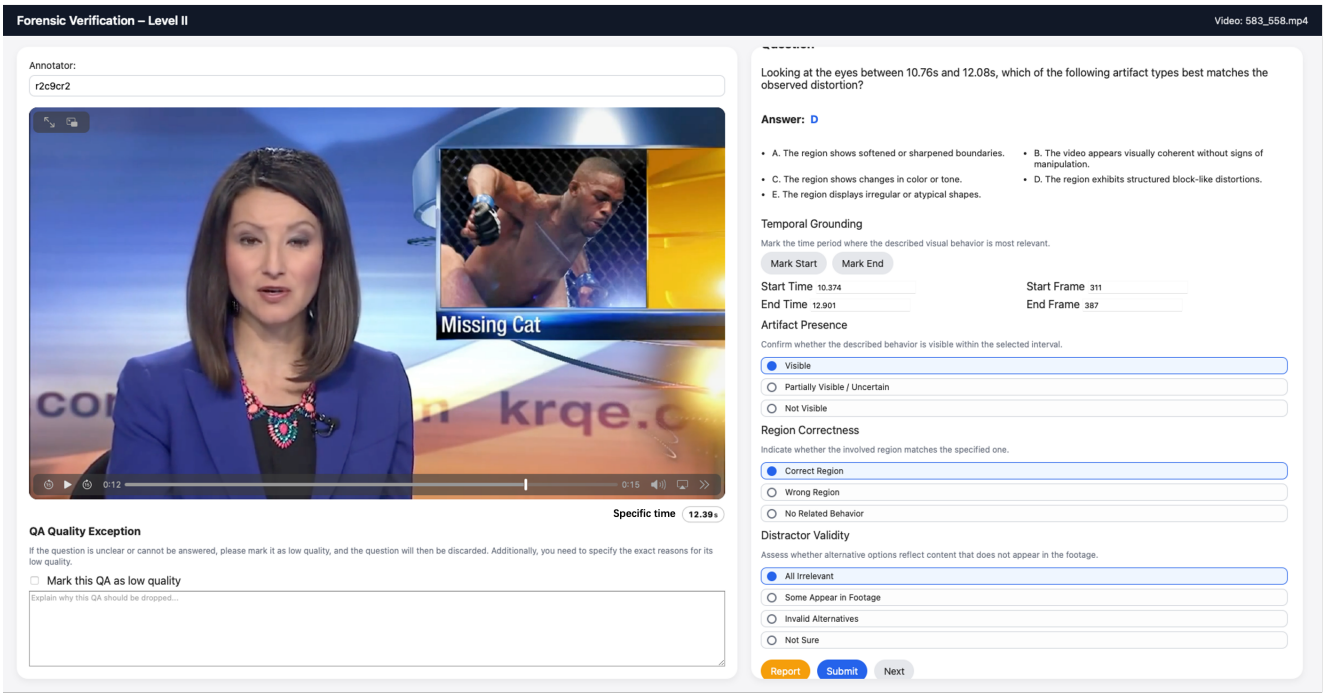


Figure 4. The verification interface for Level 2 (Temporal Deepfake Grounding). In contrast to Level 1, this stage requires rigorous temporal verification. Validators must manually use the "Mark Start" and "Mark End" buttons to define the precise boundaries of the artifact, ensuring the timestamps in the generated QA match the actual video content. Additional checks for Artifact Presence, Region Correctness, and Distractor Validity are implemented to prevent ambiguous or incorrectly grounded samples.

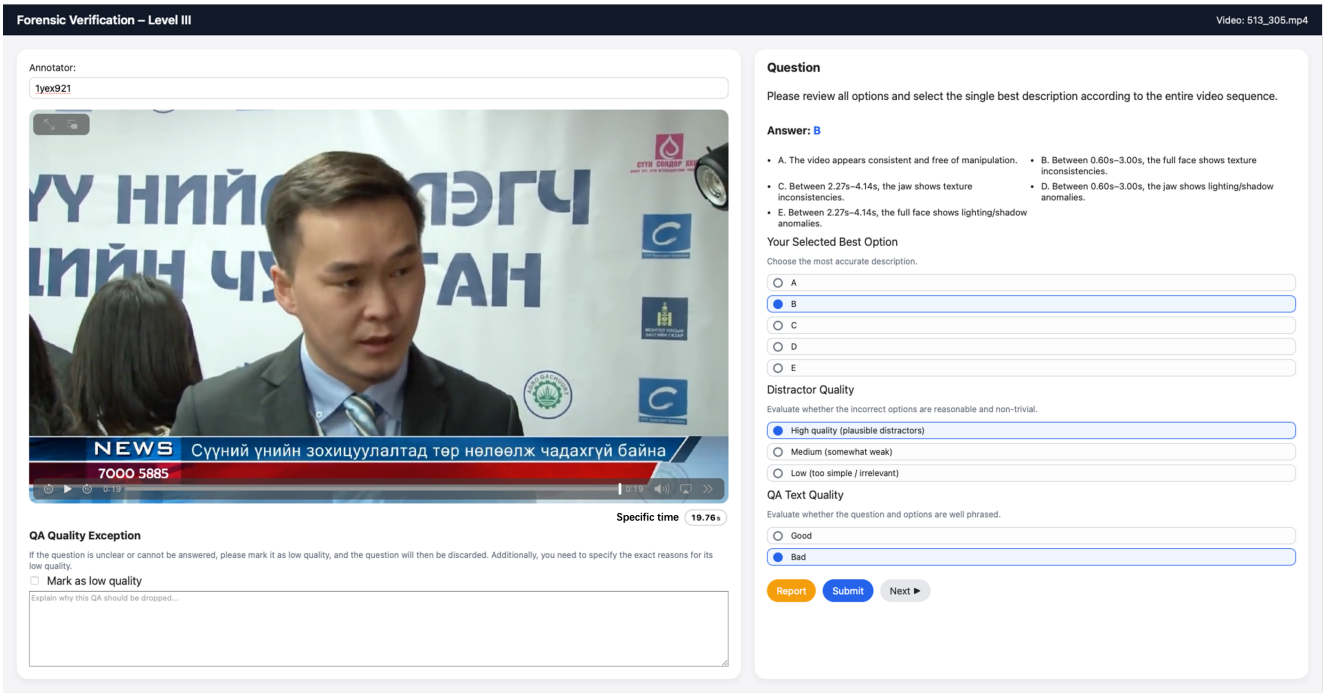


Figure 5. The verification interface for Level 3 (Forensic Reasoning). Unlike lower-level tasks, this stage requires validators to perform a holistic assessment of the entire video sequence. Validators are tasked with: (1) independently selecting the most accurate description among complex options to confirm the ground truth's validity; and (2) evaluating Distractor Quality, ensuring that incorrect options are "plausible and non-trivial" to guarantee the task effectively challenges the model's reasoning capabilities.