

# Agentic Video Summarization via Self-Reflecting Multimodal Understanding

## Supplementary Material

### 1. Ablation on the Verifier Module

To assess the contribution of the Verifier module, we conduct an ablation study on the SumMe dataset, comparing the full model with a variant where the Verifier is removed. In this ablated setting, the Reflector is forced to re-score all frames uniformly, and its outputs are directly merged with the Summarizer’s initial scores. In contrast, the full model uses the Verifier to selectively identify unreliable scores, allowing the Reflector to focus its calibration only on these ambiguous cases. As shown in Table 2, the absence of the Verifier results in consistently lower performance, demonstrating that indiscriminate re-scoring introduces unnecessary noise and weakens the corrective effect. These results confirm that the Verifier enables more precise and efficient refinement, playing a key role in improving the accuracy and stability of the final predictions.

Table 1. Ablation on the Verifier module. The study focuses on comparison between w/o Verifier (uniform re-scoring) and w/ Verifier (selective calibration).

Model	$\tau$	$\rho$
w/o Verifier	0.235	0.262
w/ Verifier	0.274	0.308

### 2. Complete Prompt Specification

#### 2.1. Verifier

We present all detailed prompts and simulated outputs for Verifier in Table 2.

#### 2.2. Reflector

We present all detailed prompts and simulated outputs for Verifier in Table 3.

### 3. Additional Visualizations.

In this section, we provide additional visualizations to further illustrate the model’s behavior as detailed in Figure 1.

Table 2. Complete prompt for Verifier.

Phase	Prompt	Simulated Output
Round 1 — Task Clarification	<p><b>SYSTEM ROLE:</b> You are an intelligent Verifier for video summarization. Your task is to evaluate initial frame-level importance scores <math>\{s_t\}</math> predicted by a Summarizer. For each frame, output a confidence value <math>c_t \in [0, 1]</math> indicating how likely the score is correct. A low confidence (<math>c_t &lt; 0.5</math>) indicates the score may be unreliable and the frame might be a missing key frame. Do not modify <math>s_t</math> directly. Remember this task and use it consistently in all subsequent rounds.</p> <p><b>USER PROMPT:</b> Confirm that you understand your task and briefly explain how you will evaluate the scores, using your memory of this instruction in future rounds.</p>	<p>I understand my task. I will evaluate each initial score <math>\{s_t\}</math> by analyzing the video content, identifying key events and frame importance. For each frame, I will assign a confidence value <math>c_t \in [0, 1]</math> representing the likelihood that the initial score is correct. Low confidence frames indicate potentially missed key frames. I will retain these instructions and apply them consistently in the following rounds.</p>
Round 2 — Learn Phase	<p><b>SYSTEM ROLE:</b> You are the Verifier and retain memory of Round 1.</p> <p><b>USER PROMPT:</b> You are given two example videos in .mp4 format with their ground truth scores <math>\{g_t\}</math>. - Video 1: <math>\langle file\_path_1 \rangle</math>, Ground truth scores: <math>\langle g_{t1} \rangle</math> - Video 2: <math>\langle file\_path_2 \rangle</math>, Ground truth scores: <math>\langle g_{t2} \rangle</math> Task: 1. Analyze each video directly to understand its content and frame-level events. 2. Internally generate video-level captions and frame-level descriptions as needed. 3. Learn scoring patterns from the ground truth, focusing on event transitions, semantic changes, and motion intensity. 4. Remember these scoring rules for the next round. Output: - A summary of the scoring rules and patterns learned from the examples.</p>	<p>From the two example videos, I learned the following scoring patterns:</p> <ol style="list-style-type: none"> <li>1. Frames containing significant scene changes or action transitions are usually assigned high scores.</li> <li>2. Continuous low-activity frames or repeated content are assigned low scores.</li> <li>3. Motion intensity and semantic novelty contribute to importance.</li> </ol> <p>I will remember these rules to guide evaluation of new videos in the next round.</p>
Round 3 — Play Phase	<p><b>SYSTEM ROLE:</b> You are the Verifier with memory from Rounds 1–2.</p> <p><b>USER PROMPT:</b> Evaluate the following test video: - Video file: <math>\langle test\_video.mp4 \rangle</math> - Initial scores <math>\{s_t\}</math>: <math>\langle</math>from Summarizer<math>\rangle</math> Task: 1. Analyze the video directly to understand content and frame-level events. 2. Internally generate captions and frame descriptions as needed. 3. Using the scoring rules learned in Round 2, evaluate each frame’s initial score. 4. For each frame, output a confidence value <math>c_t \in [0, 1]</math> and a brief reasoning. 5. Use memory of previous rounds for consistent, context-aware evaluation. 6. Do not modify <math>s_t</math>; only provide confidence values.</p>	<p>Frame 1: <math>c=0.80</math> (intro scene, low activity, initial score reasonable) Frame 2: <math>c=0.30</math> (significant action, initial score too low <math>\rightarrow</math> possible missed key frame) Frame 3: <math>c=0.75</math> (scene transition captured, score reasonable) Frame 4: <math>c=0.25</math> (important object appears, low initial score indicates possible missing key frame) ...</p>

Continued on next page

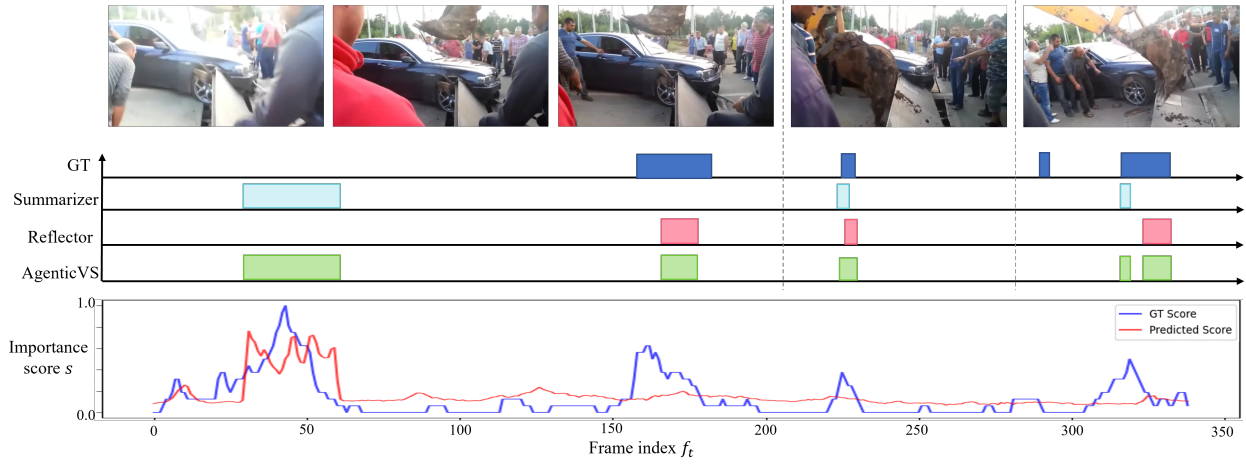
Phase	Prompt	Simulated Output
Round 4 — Confidence Analysis	<p><b>SYSTEM ROLE:</b> You are the Verifier with memory of Rounds 1–3, including initial scores <math>\{s_t\}</math> and confidence values <math>\{c_t\}</math>.</p> <p><b>USER PROMPT:</b> 1. Determine an adaptive low-score threshold <math>\theta_s</math> for the test video: <math>\theta_s = m(s_t) - 0.5 * \sigma(s_t)</math> 2. Identify frames where <math>s_t &lt; \theta_s</math> AND <math>c_t &lt; 0.5</math>. 3. List the indices of these frames and briefly explain why the Summarizer may have missed them, using learned scoring rules.</p>	<p>Adaptive threshold <math>\theta_s</math> calculated: 0.45 Frames meeting <math>s_t &lt; \theta_s</math> AND <math>c_t &lt; 0.5</math>: [2,4] Explanation: - Frame 2: significant action detected, initial score underestimated → likely missed key frame - Frame 4: appearance of important object, low initial score → missed key frame</p>
Round 5 — Final Verification	<p><b>SYSTEM ROLE:</b> You are the Verifier with full context from previous rounds.</p> <p><b>USER PROMPT:</b> Summarize your final reasoning for the video: - Frames requiring re-evaluation - Reason for unreliable initial scores - Global confidence for the video Output format: Re-evaluation frames: <math>[t_1, t_2, \dots]</math> Reason: ... Global confidence: 0.xx</p>	<p>Re-evaluation frames: [2,4] Reason: Initial scores for these frames underestimated their importance due to significant actions and object appearances. Summarizer did not fully capture inter-frame semantic transitions. Global confidence: 0.68</p>

Table 3. Complete prompt for Reflector.

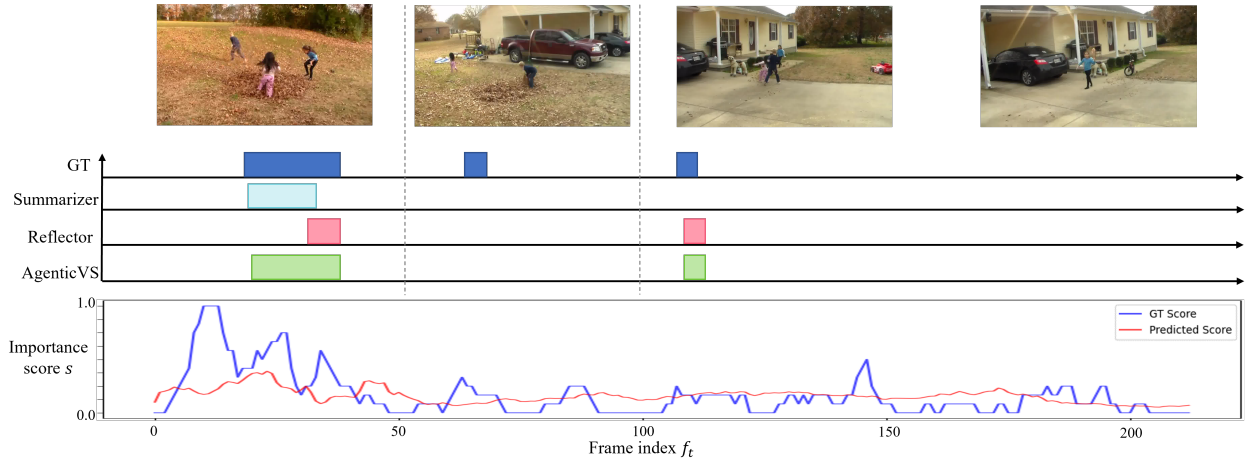
Phase	Prompt	Simulated Output
Round 1 — Task Clarification	<p><b>SYSTEM ROLE:</b> You are an intelligent Reflector for video summarization. Your task is to autonomously analyze video content and re-evaluate frames flagged by the Verifier for recalibration. For each frame, generate captions representing what content should be included in the summary. These captions will later be paired with the corresponding frames and input to CLIP to compute calibrated importance scores. Do not provide final scores directly. Remember this task and use it consistently in all subsequent rounds.</p> <p><b>USER PROMPT:</b> Confirm that you understand your task and briefly explain how you will generate captions and guide the recalibration process using your memory of this instruction in future rounds.</p>	<p>I understand my task. I will analyze video content and generate captions summarizing key content, focusing on overall events, scene changes, and transitions. I will retain these instructions for consistent use in subsequent rounds, preparing the data for calibrated scoring without directly providing scores.</p>

Continued on next page

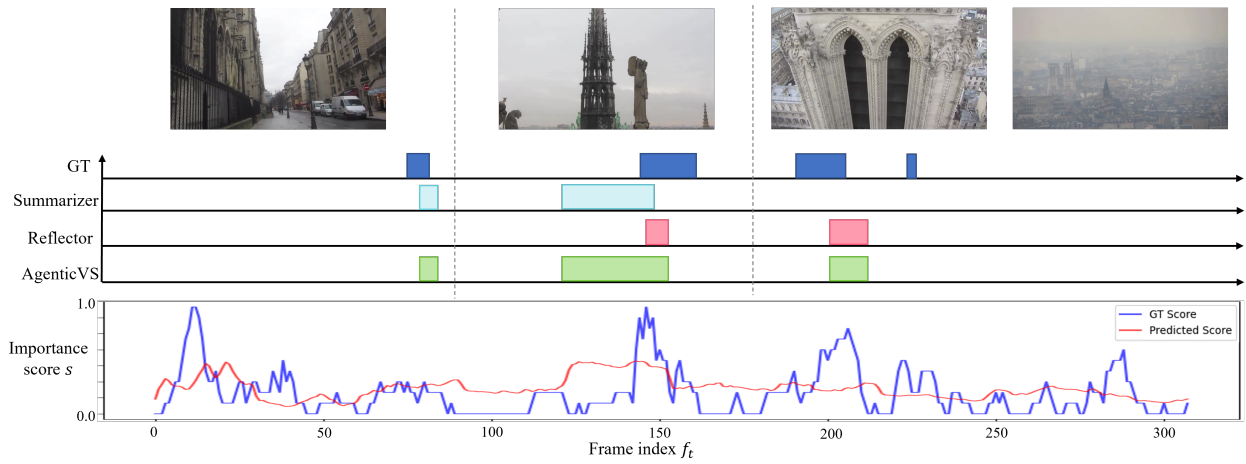
Phase	Prompt	Simulated Output
Round 2 — Video Analysis & Caption Generation	<p><b>SYSTEM ROLE:</b> You are the Reflector with memory from Round 1.</p> <p><b>USER PROMPT:</b> You are given a video file: &lt;video.mp4&gt;.</p> <p><b>Task:</b> Analyze the video directly to understand overall content, event transitions, and scene changes. Generate caption <math>C</math> summarizing what the video summary should include. The caption will later be paired with frames flagged for recalibration.</p> <p><b>Output:</b> Caption <math>C</math> representing the video’s key content.</p>	<p>Caption <math>C</math>: "A person enters a room, picks up an object, and walks across the room; multiple scene transitions and camera movements highlight different activities."</p>
Round 3 — Frame Preparation for CLIP	<p><b>SYSTEM ROLE:</b> You are the Reflector with memory from Rounds 1–2.</p> <p><b>USER PROMPT:</b> Frames requiring recalibration are: indices from Verifier.</p> <p><b>Task:</b> Extract the corresponding frame images <math>\{\hat{f}_t\}</math> from &lt;video.mp4&gt;. Pair each frame <math>\{\hat{f}_t\}</math> with the caption <math>C</math> from Round 2. Prepare the paired data for CLIP-based scoring.</p> <p><b>Output format:</b> List of frame-caption pairs ready for CLIP input.</p>	<p>Frame-caption pairs: Frame 2: paired with Caption <math>C</math> Frame 4: paired with Caption <math>C</math></p>
Round 4 — CLIP Scoring	<p><b>SYSTEM ROLE:</b> You are the Reflector with memory from Rounds 1–3, including frame-caption pairs.</p> <p><b>USER PROMPT:</b></p> <p><b>Task:</b> Feed each frame-caption pair into CLIP and compute cosine similarity. Output the calibrated scores <math>\{\hat{s}_t\}</math> for each frame, maintaining correspondence with <math>\{\hat{f}_t\}</math></p>	<p>Frame 2: <math>\hat{s}_2=0.78</math> Frame 4: <math>\hat{s}_4=0.82</math></p>
Round 5 — Score Normalization and Replacement	<p><b>SYSTEM ROLE:</b> You are the Reflector with full context from previous rounds, including <math>\{\hat{s}_t\}</math> and initial scores <math>\{s_t\}</math>.</p> <p><b>USER PROMPT:</b></p> <p><b>Task:</b> Normalize both initial scores <math>\{s_t\}</math> and CLIP-calibrated scores <math>\{\hat{s}_t\}</math>. Rescale <math>\{\hat{s}_t\}</math> to match the scale of <math>\{s_t\}</math>. Replace inaccurate initial scores with the rescaled calibrated scores. Re-evaluate key frame positions based on the updated importance scores and output the final video summarization result.</p>	<p>Frame 2: calibrated score = 0.77 Frame 4: calibrated score = 0.81 Final key frame positions: [2,4] Summary: Calibrated scores replace underestimated initial scores. Frames re-evaluated based on global video context, capturing previously missed key content.</p>



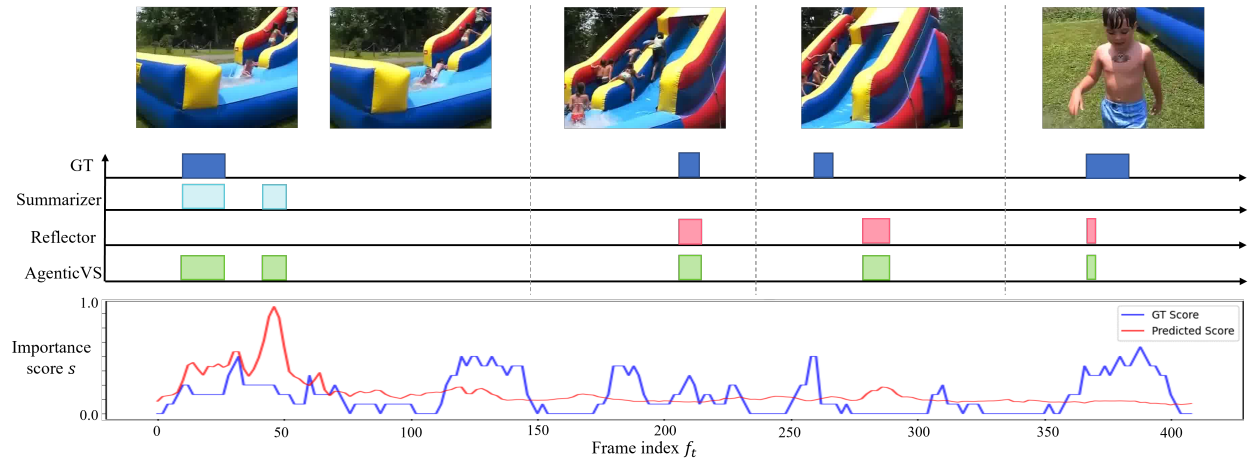
(1) Video 6. Car\_railcrossing



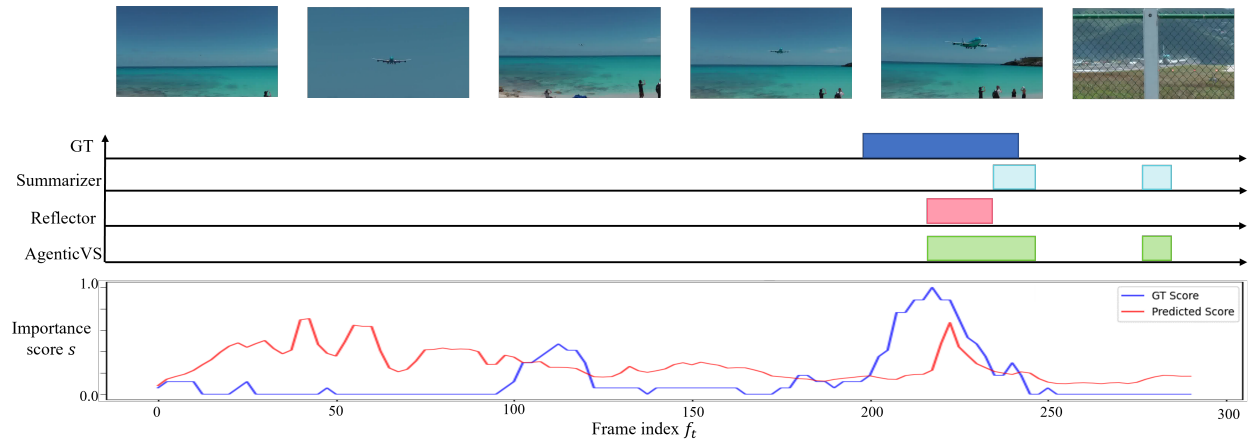
(2) Video 13. Kids\_playing\_in\_leaves



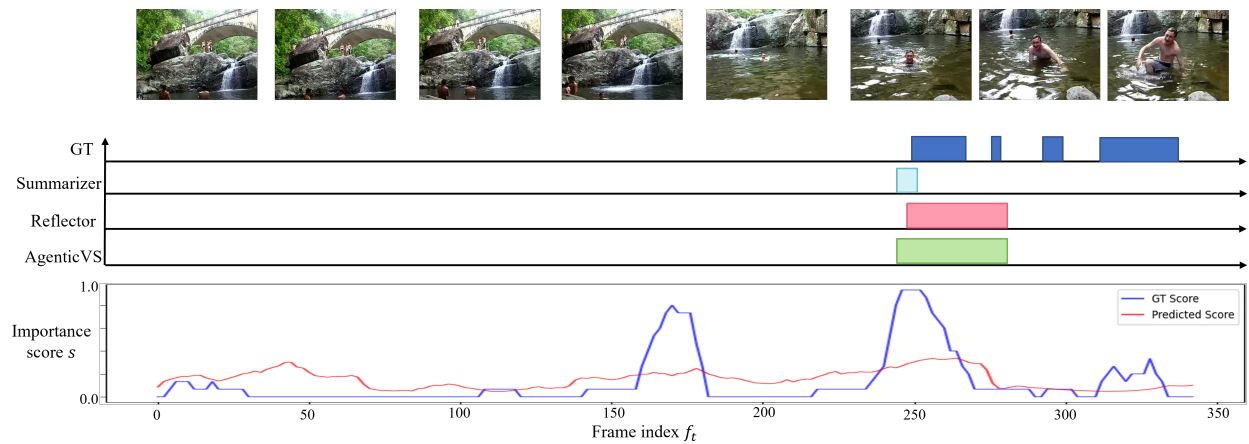
(3) Video 14. Notre\_Dame



(4) Video 16. Playing\_on\_water\_slide



(5) Video 19. St Maarten Landing



(6) Video 24. paluma\_jump

Figure 1. Additional visualization results.