A. Details of LongVR Dataset.

We present a detailed breakdown of the data distribution and key statistics for LongVR dataset in Table 8. Specifically, ActivityNet-Caption contains 37,421 instances with an average caption length of 15.80 words and an average video duration of 171.44 seconds. OpenVid-1M comprises 30,000 instances, featuring longer captions with an average length of 143.16 words and shorter videos averaging 9.07 seconds. ViTT has 5,086 instances, with captions averaging 24.98 words and videos lasting 305.08 seconds on average. MovieChat-Caption includes 808 instances, having the longest captions at an average of 143.47 words, and videos with an average duration of 457.65 seconds. Youcook2 consists of 8,700 instances, with captions averaging 80.90 words and videos lasting 383.19 seconds. Collectively, the LongVR dataset encompasses a total of 82,015 instances, with an overall average caption length of 71.12 words and an average video duration of 145.62 seconds. This diverse and extensive dataset enables robust training of the ASP-CLIP model, capturing a wide range of video and caption characteristics.

Dataset	Instance Num	Avg Caption Length	Avg Video Length
ActivityNet-Caption	37,421	15.80	171.44
OpenVid-1M	30,000	143.16	9.07
ViTT	5,086	24.98	305.08
MovieChat-Caption	808	143.47	457.65
Youcook2	8,700	80.90	383.19
Total	**82,015**	**71.12**	**145.62**

Table 8. Instance numbers of different datasets for training the ASP-CLIP model.

B. More Experiment Details and Results.

B.1. Model Selection for Different Tasks

For VideoMME, we employ InternVL2.5 (78B & 8B) and Qwen2VL (72B). For EgoSchema, we use Qwen2VL (72B & 7B) and LLaVAVideo (72B). For MLVU and LongVideoBench, we utilize LLaVAVideo (72B) and InternVL2.5 (78B & 8B). Our selection mechanism chooses different agent groups based on pseudo labels and dataset-specific features.

B.2. Efficiency Analysis

We evaluate a total of 13,943 videos across the datasets LongVideoBench, MLVU, VideoMME (with and without subtitles), and EgoSchema. For comparison, Qwen2VL (72B) requires 350.5 hours in total, averaging 90.5 seconds per video. We break down the time consumption of our approach into the following steps: Fine-tuning, conducted with 95M parameters, takes 3 hours. Preselection on 750 videos consumes 6.58 hours, averaging 31.6 seconds per video. The evaluation phase takes 130.13 hours, with an average of 33.6 seconds per video. In total, the evaluation process (including preselection and evaluation) requires 139.71 hours, averaging 36.07 seconds per video—this is 2.5× faster than Qwen2VL.

B.3. The Result of Using More Than Three Agents

The performance of the 4-agent setup (EgoSchema: 83.0, LongVideoBench: 80.1, MLVU: 84.1, VideoMME: 81.8 / 86.5) is slightly superior to that of the 3-agent configuration, but it incurs additional time cost—taking 38.7 seconds per video, which is 5.1 seconds longer than the 3-agent setup. Considering the trade-off between accuracy and efficiency, we opt for the 3-agent configuration. Notably, LVAgent can be easily extended to incorporate more agents if needed.

C. Prompts of LVAgent.

In this section, we present the comprehensive set of prompts utilized by LVAgent across various stages of its operation, as shown in Table 9. These prompts are meticulously designed to guide LVAgent in effectively performing its tasks, ensuring seamless interactions and accurate responses.

Prompt	Content
Prompt to decide whether to watch the whole video	You are given a single-choice question, options, subtitles, and some frames of the long video. You should not only look at the textual information but also consider the input visual information, taking everything into account. If you can answer the question accurately and comprehensively based on the existing information, especially the visual information, and further watching the entire video will not significantly improve the quality of the answer, then you don't need to watch the entire video and can answer 'No.'. However, if the existing information is not sufficient to fully answer the question, and watching the entire video may obtain information crucial for answering the question, please reply 'Yes' The frame tokens: {`Frame tokens`} {`Question`} {`Options`} {`Subtitiles`}(If have) Output:{`Yes/No`}
Prompt for generate the key information	Given four randomly sampled frames from a long video, subtitles, a question, and multiple-choice options, identify the key information needed to answer the question. Focus on visual cues, context, and temporal relationships within the frames. Limit your response to 50 words. The frame tokens: {`Frame tokens`} {`Question`} {`Options`} {`Subtitles`}(If have)
Prompt for generating the answer	Select the best answer to the following multiple-choice question based on the video and the subtitles. Respond with only the letter (A, B, C, or D) of the correct option. The frame tokens: {`Frame tokens`} {`Question`} {`Options`} {`Subtitles`} (If have) The best answer is:
Prompt for generating the reason	Given the video frames you've seen, and the question along with your answer, deeply analyze the logical steps and evidence from the frames that led you to provide this particular answer. The Question is: {`Question`} The predict answer is {`Predict answer`}

Prompt	**Content**
Prompt for generating scores for agents	You are given the answers and the reasoning for judgment from this model and two other models. The question is: {`Question`} The answer of {`Agent 1`} is {`Agent 1's Answer`} the reason is {`Agent 1's Reason`} The answer of {`Agent 2`} is {`Agent 2's Answer`} the reason is {`Agent 2's Reason`} The answer of {`Agent 3`} is {`Agent 3's Answer`} the reason is {`Agent 3's Reason`} Please score the performance of these three agents based on their reasoning. The score ranges from 1 to 10. Please strictly follow the answer formatThe answer format is: {`Agent 1's Score`}: 1-10 {`Agent 2's Score`}: 1-10 {`Agent 3's Score`}: 1-10 The reason is: {`Reason`}
Prompt for history information summarization	Agent 1's answer is: {`Agent 1's Answer`}. Reason: {`Agent 1's Reason`}. The final score is : {`Agent 1's Score`}. Agent 2's answer is: {`Agent 2's Answer`}. Reason: {`Agent 2's Reason`}. The final score is : {`Agent 2's Score`}. Removed {`Agent 3`} Answer Answer: {`Agent 3's Answer`} Reason {`Agent 2's Reason`} However, this reason was deemed unconvincing, so this answer was removed from the discussion.

Table 9. The prompting templates used in different key steps of LVAgent.