VideoAutoArena: An Automated Arena for Evaluating Large Multimodal Models in Video Analysis through User Simulation

Supplementary Material

A. Prompts

A.1. User Simulation

In Figure 1 and 2, we include the prompts for video content-constrained persona generation and persona-constrained video question asking.

A.2. Fault-Driven Evolution

Figure 3 includes the prompt for the fault-driven evolution.

A.3. Automatic Judging

Figure 4 includes the prompt for the automatic judging. Our VideoAutoBench adopts the same prompt for judging.

A.4. Difficulty Level Evaluation

Figure 5 includes the prompt for the complexity evaluation.

B. Experimental Details

B.1. Statistics

In VideoAutoArena, Model A wins 5,620 battles, Model B wins 5,941 battles, and 918 ties. VideoAutoBench consists of 255 samples corresponding to 244 unique videos, with an average duration of 478.5 seconds. The duration distribution includes 62 videos for 8-15s, 62 for 15-60s, 60 for 180-600s, and 60 for 900-3,600s. The samples span 10 categories: 29 from Movies, 50 from Life Vlogs, 9 from Geography, 13 from History, 12 from News Programs, 9 from Art, 6 from STEM, 8 from Computer Science, 55 from Cooking Recipes, and 53 from Travel Guides.

B.2. LMMs Selection

In our experiments, we evaluate 11 SOTA LMMs:

1. `gpt-4o-2024-05-13`,
2. `gpt-4o-mini-2024-07-18`,
3. `gemini-1.5-pro`,
4. `gemini-1.5-flash`,
5. `rhymes-ai/Aria`,
6. `Qwen/Qwen2-VL-72B-Instruct`,
7. `Qwen/Qwen2-VL-7B-Instruct`
8. `lmms-lab/LLaVA-Video-72B-Qwen2`,
9. `lmms-lab/LLaVA-Video-7B-Qwen2`,
10. `lmms-lab/llava-onevision-qwen2-72b-ov`,
11. `lmms-lab/llava-onevision-qwen2-7b-ov`.

We selected the top 10 LMMs that support long video analysis, along with their smaller versions, based on their performance on VideoMME as of October 15, 2024. Additional open-source LMMs were excluded for two reasons: first, some were released concurrently with our work, leaving insufficient time for evaluation; second, others exhibited weak performance on VideoMME or lacked support for long video analysis. Consequently, our experiments are limited to the 11 most popular and SOTA LMMs.

For user simulation, fault-driven evolution, automatic judging, and difficulty level evaluation, we use `gpt-4o-2024-08-06`, ensuring that the examiner and judge remain distinct from the LMMs throughout the entire evaluation process.

B.3. Hyperparameters

For user simulation, fault-driven evolution, automatic judging, and difficulty level evaluation, each video is uniformly sampled into a maximum of 128 frames, while response generation uses up to 64 frames. Each frame is resized to 512×512. For API-based LMMs, the `max_tokens` parameter is set to 4096, with other settings using default values. For open-source LMMs, the temperature is set to 0.7, and `max_new_tokens` is limited to 2048.

C. Examples

C.1. Persona

As shown in Figure 6, we include 15 examples of different personas for 5 different videos.

C.2. Question

As shown in Figure 6, we include 15 examples of different questions for 5 different videos.

C.3. Responses and Judging

As shown in Figure 7, 8, 9, and 10, we include the battle examples between different models.

D. Human Annotations

D.1. Question Ranking

In Figure 11, we include our guideline for the question ranking annotation.

D.2. Judging

In Figure 12, we include our guideline for the response judging annotation.

````
You are watching a video. Your task is to create three distinct user personas who might be interested in this video:

- Persona 1: A user whose background is **closely related** to the video's content.
- Persona 2: A user whose background is **less related** but who may still develop an interest in learning more about the video's content.
- Persona 3: A user whose background is **totally not related** but who may still develop an interest in learning more about the video's content.

Keep each persona description brief, using only a short paragraph. Please adhere to the following format:

P1 [Persona 1]: A person xxx
P2 [Persona 2]: A person xxx
P3 [Persona 3]: A person xxx
````

Figure 1. The prompt for video content-constrained persona generation.

~~~
You are given a list of image frames from a video, and you **MUST** pretend that you are watching the video. You will act as a user with the following persona:
```persona
{persona}
```

Based on your persona, your task is to craft high-quality [Questions] for Video Question Answering (Video QA). You are required to watch a video.

Your [Question] should be based on the video and seek assistance from the AI to analyze its content. **The [Question] must reflect your persona.**

To create the [Question], please follow the steps below:

Step 1: Watch the video carefully and craft a [Question] that aligns with the user's persona. The [Question] should only be answerable by someone who has watched the video. You **CANNOT** reveal any key visual details in the [Question]. If necessary, include response format constraints (Markdown, JSON, Table, List, etc.) in the [Question].

Step 2: Watch the video carefully and provide a [Response] to the [Question]. The [Response] should be clear and organized, fully addressing the persona's needs without omitting important information. It should provide a targeted answer based on the user's perspective.

Please reply strictly in the following format:

Step 1 [Question]:
Step 2 [Response]:
~~~

Figure 2. The prompt for persona-constrained video question asking.

E. Limitation

VideoAutoArena and VideoAutoBench currently lack evaluations for multi-turn and non-English interactions, primarily due to the limited multi-turn conversational capabilities and restricted non-English proficiency of current open-sourced LMMs. Moreover, the automatic judging system tends to favor detailed responses, a preference also observed in human evaluations. While detailed responses are often more helpful to users, this introduces challenges in evaluating LMMs. We implemented the style-control method from LMSYS Chatbot Arena to adjust ELO ratings by penalizing stylistic factors. However, we found this approach unsuitable for evaluating current LMMs. For example, while Aria and Qwen2-VL-72B outperform Gemini-1.5-Pro in ELO ratings, Gemini-1.5-Pro consistently achieves significantly higher win rates. Manual review of Gemini-1.5-Pro's outputs revealed that its responses are not only more detailed but also of higher quality compared to those from the two open-source LMMs. Additionally, most open-source LMMs tend to generate less detailed responses than proprietary LMMs, causing the style-control mechanism to disproportionately penalize proprietary models. This imbalance leads to unfair evaluations. To address this issue, a more effective style-control mechanism should ensure that competing LMMs produce responses with comparable levels of detail, thereby enabling a fairer evaluation.

To address these limitations, future work will focus on expanding VideoAutoArena and VideoAutoBench to include multiturn and multilingual data. Additionally, we aim to refine our definitions of user simulation, developing a systematic approach for generating battles that encompass a broader and more inclusive range of scenarios while maintaining high separability and alignment with human judgment. Furthermore, we plan to explore advanced style-control techniques and unbiased LMMs-as-judge to further enhance the robustness and fairness of our LMM-based evaluation framework.

~~~
**Remember: You are watching a Video.**

A user, characterized by a specific persona, is interacting with two AI assistant models (A and B) to better understand video content using the same question. Here is the user's persona:
```persona
{persona}
```

The user's question is:
```question
{question}
```

The response from Model A is:
```model_a
{answer_a}
```

The response from Model B is:
```model_b
{answer_b}
```

Your task is to carefully evaluate the responses of Model A and Model B to identify any faults and models' weaknesses. Based on these weaknesses, generate a harder [New Question] that explores aspects where the models may struggle. **Remember**, the [New Question] should continue to align with the user's persona, without necessarily being more specific.

You **MUST** follow these steps to generate the [New Question]:

Step 1: Carefully review the chat history to identify any problems in the two models' responses related to understanding the video's content. Consider:
- Whether the responses correctly incorporate information from the video.
- Whether the responses are helpful in fulfilling the user's requirements.
- Whether the responses are user-aware and align with the user's persona.
List all identified faults in [Fault List A] for Model A and [Fault List B] for Model B.

Step 2: Based on [Fault List A] and [Fault List B], summarize the weaknesses of Model A and Model B. List all identified weaknesses in [Weakness List A] for Model A and [Weakness List B] for Model B. The weaknesses should be more general than the faults.

Step 3: Based on [Weakness List A] and [Weakness List B], craft a high-quality harder [New Question] that will further explore aspects where the two models may struggle. If both responses do not contain any faults, you still need to create a different harder [New Question] that mimics the interaction between the user and the AI assistant models. The [New Question] is **not** a follow-up question.

**Remember**, the [New Question] should continue to align with the user's persona, without necessarily being more specific. Ensure that:
- The [New Question] can only be answered by someone who has watched the video.
- Do **NOT** leak key visual details in the [New Question].
If necessary, include response format requirements (Markdown, JSON, Table, List, etc.) in the [New Question].

Please respond strictly in this format:

Step 1:
```[Fault List A]
xxx
```

```[Fault List B]
xxx
```

Step 2:
```[Weakness List A]
xxx
```

```[Weakness List B]
xxx
```

Step 3:
```[New Question]
Only include the new question here!
```
~~~

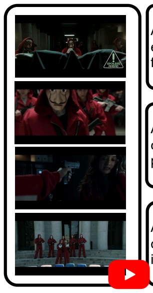
Figure 3. The prompt for our fault-driven evolution generates new questions based on the responses from the two models.

~~~
**Remember: You are watching a Video.**

A user, characterized by a specific persona, is interacting with two AI assistant models (A and B) to better understand video content using the same question. Here is the user's persona:
```persona
{persona}
```

The user's question is:
```question
{question}
```

The response from Model A is:
```model_a
{answer_a}
```

The response from Model B is:
```model_b
{answer_b}
```

Please act as an impartial judge and carefully evaluate the responses of Model A and Model B to determine which one is better. Use the following standards:

1. [Instruction Following]: The response should closely adhere to the user's instructions, ensuring it directly addresses the specified task.
2. [Accuracy]: The response must accurately utilize information from the video, avoiding fabrication or misquotation. It should maintain factual correctness, avoid hallucinations, and demonstrate contextual coherence with precise terminology and knowledge.
3. [Relevance]: The response should consider the user's background information and needs, providing a comprehensive, detailed answer that addresses the question directly without straying off-topic. Responses should be thorough, offering multiple perspectives where relevant.
4. [Helpfulness]: The response should provide valuable information to aid the user in understanding or solving their issue, avoiding irrelevant or vague content.

If the responses from Model A and Model B are of similar quality (whether both are good or both are bad), you may declare a tie.

**Please follow these steps for your judgment:**

- Step 1: Analyze which model provides a better response for the [Instruction Following] standard.
- Step 2: Analyze which model provides a better response for the [Accuracy] standard.
- Step 3: Analyze which model provides a better response for the [Relevance] standard.
- Step 4: Analyze which model provides a better response for the [Helpfulness] standard.
- Step 5: Based on the results from Steps 1-4, determine the overall outcome: Model A, Model B, Tie (both are good), or Tie (both are bad).

Please respond strictly in the following format:

```[Instruction Following]
[Your Analysis]
```

```[Accuracy]
[Your Analysis]
```

```[Relevance]
[Your Analysis]
```

```[Helpfulness]
[Your Analysis]
```

```[Overall Judge]
A/B/Tie
```
~~~

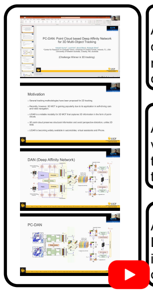
Figure 4. The prompt for our automatic judging.

~~~
**Remember: You are watching a Video.**

A user, characterized by a specific persona, is seeking help from an AI assistant to better understand video content. Here is the user's persona:
```persona
{persona}
```

The first question is:
```Question A
{question_a}
```

The second question is:
```Question B
{question_b}
```

Please act as an impartial evaluator and assess the difficulty level of these two questions based on the following criteria (1 = easiest, 5 = hardest):

1. [Instruction Following]: Rate the complexity (1-5) involved in accurately following the question’s instructions.
2. [Accuracy]: Rate the challenge (1-5) in using specific information from the video to address the question without errors or fabrications.
3. [Relevance]: Rate the difficulty (1-5) of tailoring the response to the user's background and needs, ensuring alignment with the context.
4. [Helpfulness]: Rate the level of difficulty (1-5) in providing information that is genuinely valuable and directly aids the user, avoiding vague or irrelevant content."

Please respond strictly in the following format:

1. Question A
```[Question A Instruction Following Difficulty Level]
[Your Score]
```

```[Question A Accuracy Difficulty Level]
[Your Score]
```

```[Question A Relevance Difficulty Level]
[Your Score]
```

```[Question A Helpfulness Difficulty Level]
[Your Score]
```

```[Question A Overall Difficulty Level]
[Overall Difficulty Score 1-5]
```

2. Question B

```[Question B Instruction Following Difficulty Level]
[Your Score]
```

```[Question B Accuracy Difficulty Level]
[Your Score]
```

```[Question B Relevance Difficulty Level]
[Your Score]
```

```[Question B Helpfulness Difficulty Level]
[Your Score]
```

```[Question B Overall Difficulty Level]
[Overall Difficulty Score 1-5]
```
~~~

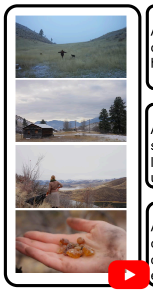
Figure 5. The prompt for question complexity evaluation.



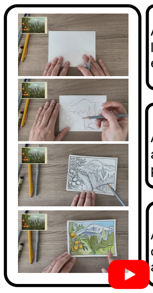
Persona 1 A person who is a fan of heist dramas and has watched "Money Heist" multiple times. They enjoy analyzing plot twists and character development, and they are always on the lookout for video recaps or discussions to deepen their understanding of the show.	**Question 1** In the video recap of "Money Heist," how does the Professor ensure that the police and public sympathize with the robbers' cause, and what strategies are highlighted to demonstrate the calculated nature of the heist plan? Include character interactions or decisions that emphasize these strategies.
Persona 2 A person who enjoys crime series and thrillers but hasn't seen "Money Heist" yet. They are curious about popular shows and are looking for recommendations. This recap video might pique their interest and prompt them to start watching the series.	**Question 2** What was the Professor's strategy in executing the heist in "Money Heist," and how did it involve the team and hostages?
Persona 3 A person who primarily watches comedies and documentaries. They generally avoid intense dramas, but a friend recommended "Money Heist" as an exception. This video could introduce them to the show's narrative and possibly lead them to explore a new genre.	**Question 3** Considering I usually enjoy comedies and documentaries, could you tell me how the initial episodes of "Money Heist" set up the heist and the team dynamics without getting too intense or dramatic? I'm curious about the plan and the characters but want to know if it aligns with my taste.



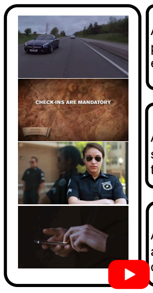
Persona 1 A person with a background in computer science, specializing in computer vision and 3D object tracking. They are actively engaged in research related to LiDAR technology and multi-object tracking systems, seeking to implement state-of-the-art techniques like point cloud-based tracking in their projects.	**Question 1** Based on the video presentation, what are the key contributions and results of the PC-DAN (Point Cloud Deep Affinity Network) in 3D multi-object tracking, especially concerning the datasets mentioned? Please summarize in a list format.
Persona 2 A person with a background in robotics engineering, working primarily on autonomous vehicles. Although their expertise is more focused on mechanical and systems engineering, they are keen to expand their knowledge in computer vision and LiDAR-integrated systems to enhance navigation and tracking capabilities in robotics.	**Question 2** How does the PC-DAN method for 3D multi-object tracking utilize point clouds, and what advantages does it offer for autonomous vehicle navigation systems compared to traditional 2D tracking methodologies?
Persona 3 A person with a background in business management, currently working in the tech industry. Despite having no technical expertise in computer vision or LiDAR technologies, they are interested in learning about cutting-edge advancements in technology to identify potential opportunities for business expansion and innovation.	**Question 3** Based on the presentation, how does the PC-DAN model for 3D multi-object tracking compare in performance to other models when tested on the JRDB dataset? Can you highlight any key business opportunities or innovations that might arise from its application in real-world scenarios?



Persona 1 A person who lives in a rural area, enjoys simple living, and has a deep interest in traditional crafts, herbalism, and folklore. They appreciate a slow-paced lifestyle and enjoy making homemade tea blends and exploring stories from different cultures.	**Question 1** How does the video illustrate the experience of exploring and sourcing herbal ingredients for tea blends and the connection to folklore and personal stories during the winter season?
Persona 2 A person who works in an urban setting but is enthusiastic about transitioning to a more sustainable lifestyle. They are curious about connecting with nature and finding new hobbies like trying cultural recipes and learning about myths, even though it's different from their usual routine.	**Question 2** How does the video illustrate the process of connecting with nature and embracing cultural traditions through activities like tea making and cooking, and what impacts do these activities have on personal well-being according to the narrator?
Persona 3 A person working in a fast-paced corporate job, typically focused on technology and modern conveniences. They stumble upon this video, becoming intrigued by the idea of slowing down and appreciating nature, new cultures, and more traditional ways of living as an escape from their daily hustle.	**Question 3** How does the person in the video incorporate elements of nature and traditional living into their daily routine, and what impact does this have on their sense of balance and well-being? I'm curious to explore how stepping away from a technology-driven lifestyle can lead to a more fulfilling life. Please provide the response in a detailed list format.

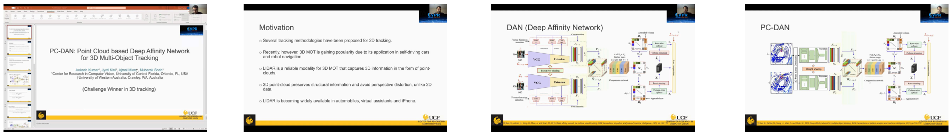


Persona 1 A person who is an art teacher looking for creative ways to teach postcard design and landscape drawing to their students. They are interested in finding engaging tutorials to enhance their class curriculum.	**Question 1** In the video, can you provide a summary of the step-by-step process shown for designing and sketching a landscape postcard? I'm interested in understanding the techniques and tools used as I plan to incorporate similar methods into my art classes.
Persona 2 A person who enjoys dabbling in various DIY projects and has a budding interest in art. They are keen to explore new crafts and may get inspired to try drawing and creating personalized postcards as a new hobby.	**Question 2** In the video, the creator is making art inspired by an old postcard. Can you describe the process they use to create their artwork and what materials are involved? How might this inspire someone interested in DIY art projects?
Persona 3 A person who primarily has a background in finance but is looking for a relaxing and creative outlet. They stumble upon this art tutorial video and find the idea of drawing and painting as a form of stress relief intriguing and worth trying out.	**Question 3** As someone new to art and looking for a creative stress relief, how did the video guide the process of drawing and painting a postcard, and what materials were used throughout the tutorial?



Persona 1 A person who is a current or former law enforcement officer with experience in witness protection programs. They have a strong interest in understanding the logistical and emotional challenges involved in relocating and protecting high-risk individuals.	**Question 1** In the context of the witness protection program depicted in the video, what logistical measures are in place to ensure the safety and integration of witnesses, and how do these measures address both security and emotional support needs?
Persona 2 A person who is a criminal justice student curious about different aspects of the legal system, including how witness protection operates. They are eager to learn beyond textbooks, especially about real-world applications like WITSEC.	**Question 2** Can you describe how the video highlights WITSEC's approach to maintaining participants' safety and their integration into new communities, as inferred from the visual elements and narration?
Persona 3 A person who is a fiction writer looking for inspiration for their next crime thriller. They seek to add realism to their storylines by exploring how witness protection programs function and the dynamics involved in protecting witnesses.	**Question 3** What elements of the witness protection program were highlighted as essential services in the video, and how do they contribute to the overall safety and acclimation of participants into new communities?

Figure 6. Examples of our user simulation include five videos from diverse domains: Movies, Computer Science, Life Vlogs, Art, and News Programs. To save space, we only showcase 4 frames of each video.



PC-DAN: Point Cloud based Deep Affinity Network for 3D Multi-Object Tracking

Aakash Kumar*, Jyoti Kini*, Ajmal Mian†, Mubarak Shah*
*Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA
†University of Western Australia, Crawley, WA, Australia

(Challenge Winner in 3D tracking)

Motivation

- Several tracking methodologies have been proposed for 2D tracking.
- Recently, however, 3D MOT is gaining popularity due to its application in self-driving cars and robot navigation.
- LIDAR is a reliable modality for 3D MOT that captures 3D information in the form of point-clouds.
- 3D point-cloud preserves structural information and avoid perspective distortion, unlike 2D data.
- LIDAR is becoming widely available in automobiles, virtual assistants and iPhone.

DAN (Deep Affinity Network)

PC-DAN

Persona: A person with a background in robotics engineering, working primarily on autonomous vehicles. Although their expertise is more focused on mechanical and systems engineering, they are keen to expand their knowledge in computer vision and LiDAR-integrated systems to enhance navigation and tracking capabilities in robotics.
Question: How does the PC-DAN method for 3D multi-object tracking utilize point clouds, and what advantages does it offer for autonomous vehicle navigation systems compared to traditional 2D tracking methodologies?

Model A: Aria

The PC-DAN (Point Cloud-based Deep Affinity Network) method for 3D multi-object tracking leverages point cloud data, which captures 3D information, to enhance tracking accuracy and robustness. This approach extends previous 2D tracking methodologies by integrating deep learning techniques to handle the complexities of 3D spatial data. The system employs a VGG network for feature extraction from point clouds and utilizes a deep affinity network to associate features across frames, ensuring accurate tracking of multiple objects.

One of the primary advantages of PC-DAN over traditional 2D tracking methods is its ability to preserve structural information and avoid perspective distortions, which are common in 2D data. This is particularly beneficial for autonomous vehicle navigation systems, as it allows for more reliable detection and tracking of objects in 3D space, leading to safer and more efficient navigation. Additionally, the use of point clouds enables the system to handle varying perspectives and occlusions more effectively, further improving performance in dynamic environments.

Model B: GPT-4o

The PC-DAN (Point Cloud based Deep Affinity Network) method for 3D multi-object tracking leverages point cloud data to enhance tracking performance in scenarios such as autonomous vehicle navigation. Here's an overview of how it utilizes point clouds and its advantages:

Utilization of Point Clouds:
1. **Feature Extraction:**
  - PC-DAN uses the PointNet architecture to process point cloud data. PointNet extracts features from the raw point cloud, capturing 3D structural information.

2. **Affinity Measurement:**
  - The extracted features from consecutive frames are used to measure the affinities between objects in 3D space. This helps in associating objects across different frames accurately.

3. **Deep Affinity Network:**
  - PC-DAN incorporates a deep affinity network that processes these features through several layers (e.g., VGG, parameter sharing, and compression network) to handle the association tasks.

4. **Consistent 3D Tracking:**
  - The network generates and maintains consistent 3D object identities by using forward and backward association losses, ensuring robust tracking performance over time.

Advantages Compared to Traditional 2D Tracking:
1. **Enhanced Structural Information:**
  - Unlike 2D data, 3D point clouds preserve detailed structural information and are not affected by perspective distortion. This makes it easier to track objects accurately in a three-dimensional space.

2. **Robustness to Occlusion:**
  - 3D data provides a more comprehensive view of the environment, reducing the impact of occlusions and ambiguities that are common with 2D data.

3. **Better Spatial Context:**
  - Point clouds offer precise spatial context and depth information, crucial for tasks like autonomous navigation where understanding the true distance and relative positions of objects is vital.

4. **Adaptability to Dynamic Environments:**
  - The use of LIDAR and point cloud data allows better adaptability and robustness in dynamic environments, such as those encountered in self-driving scenarios.

5. **Broad Application Use:**
  - LIDAR technology, which feeds point cloud data, is increasingly available in modern automobiles, enhancing the feasibility and scalability of PC-DAN-based systems for real-world applications.

Results and Validation:
- The effectiveness of PC-DAN is validated on datasets like KITTI and JackRabbit (JRDB), showing competitive performance metrics compared to other methods, with high MOTA (Multi-Object Tracking Accuracy) scores.

By utilizing these advantages, PC-DAN significantly improves the reliability and accuracy of tracking in autonomous vehicle systems, compared to traditional 2D tracking methodologies that lack depth information and are susceptible to occlusions and distortions.

Automatic Judging

Instruction Following: Both models followed the instructions well, addressing the specifics of how PC-DAN uses point clouds for 3D tracking and comparing it to 2D methods. However, Model B's response is more detailed in outlining the utilization steps and advantages explicitly. Model B is slightly better.

Accuracy: Model A is accurate but less detailed in using specific methodologies like PointNet and the incorporation of various network layers in PC-DAN. Model B accurately reflects the video content, mentioning PointNet, specific network processes, and validation on datasets. Model B is better.

Relevance: Model B provides a more thorough and technical explanation, directly catering to the user's background in autonomous vehicles and interest in LiDAR and point clouds. It also mentions validation datasets which are relevant. Model B is better.

Helpfulness: Model B offers a comprehensive explanation with distinct advantages of the method, catering well to the user's needs to understand 3D tracking systems in autonomous vehicles. It gives more specific technical insights and context. Model B is more helpful.

Model B Wins!

Figure 7. Example of the battle between Aria and GPT-4o.

Persona: A person who is an art teacher looking for creative ways to teach postcard design and landscape drawing to their students. They are interested in finding engaging tutorials to enhance their class curriculum.
Question: In the video, can you provide a summary of the step-by-step process shown for designing and sketching a landscape postcard? I'm interested in understanding the techniques and tools used as I plan to incorporate similar methods into my art classes.

Model A: GPT-4o-mini

In the video, the process for designing and sketching a landscape postcard is illustrated step-by-step. Here’s a summary of the techniques and tools used:

Step-by-Step Process:

1. **Materials Preparation**:
   - Gather tools such as various colored pencils, a fine-tip pen for outlines, erasers, and a watercolor palette. Paper is divided into sections for planning.

2. **Initial Sketch**:
   - Start by lightly drawing a basic composition on a sheet of watercolor paper. Reference images (like landscapes) can be placed nearby for inspiration.

3. **Detailed Outline**:
   - Add details to the initial sketch, refining the shapes, leaves, and any foreground elements like fruits or flowers.

4. **Adding Texture**:
   - Consider the textures of various elements (such as tree leaves and mountains) and include them in the sketches to convey depth.

5. **Incorporating Color**:
   - Use watercolor paints to fill in areas, starting with a base layer and gradually adding more vibrant colors. Pay attention to blending techniques for realism.

6. **Final Touches**:
   - Once the painting dries, outline details with a fine-tip pen to enhance the defined elements and add contrast.

7. **Personalization**:
   - Create a postcard format by adding space for a message on the back and potentially a stamp, making the artwork suitable for sending.

Techniques Used:
- **Layering**: Build up color gradually for depth.
- **Blending**: Mixing colors seamlessly on the paper.
- **Texture Creation**: Using varied strokes to reflect natural patterns.

Tools Used:
- **Watercolors**: For vibrant, fluid colors.
- **Colored Pencils**: For fine detailing and layering.
- **Fine-Tip Pens**: For outline and finishing touches.

Incorporating these methods into your art classes can encourage students to experiment with both drawing and painting techniques, enhancing their understanding of compositional elements in postcard design.

Model B: LLaVa-Vieo-72B

The video begins with a blank sheet of paper, and the artist starts by sketching the basic outline of the landscape. They use a pencil to lightly draw the horizon line, mountains, and trees. The artist then adds more details such as branches and leaves, using a variety of strokes to create texture and depth. After the sketch is complete, the artist uses watercolors to add color to the scene, starting with the sky and moving on to the trees and mountains. The final step involves adding a personal message and address to the postcard, turning it into a functional piece of art.

Automatic Judging

Instruction Following: Model A provides a detailed step-by-step process as requested, while Model B gives a more general overview. Model A follows the instructions more closely by breaking down the process into specific steps and techniques.

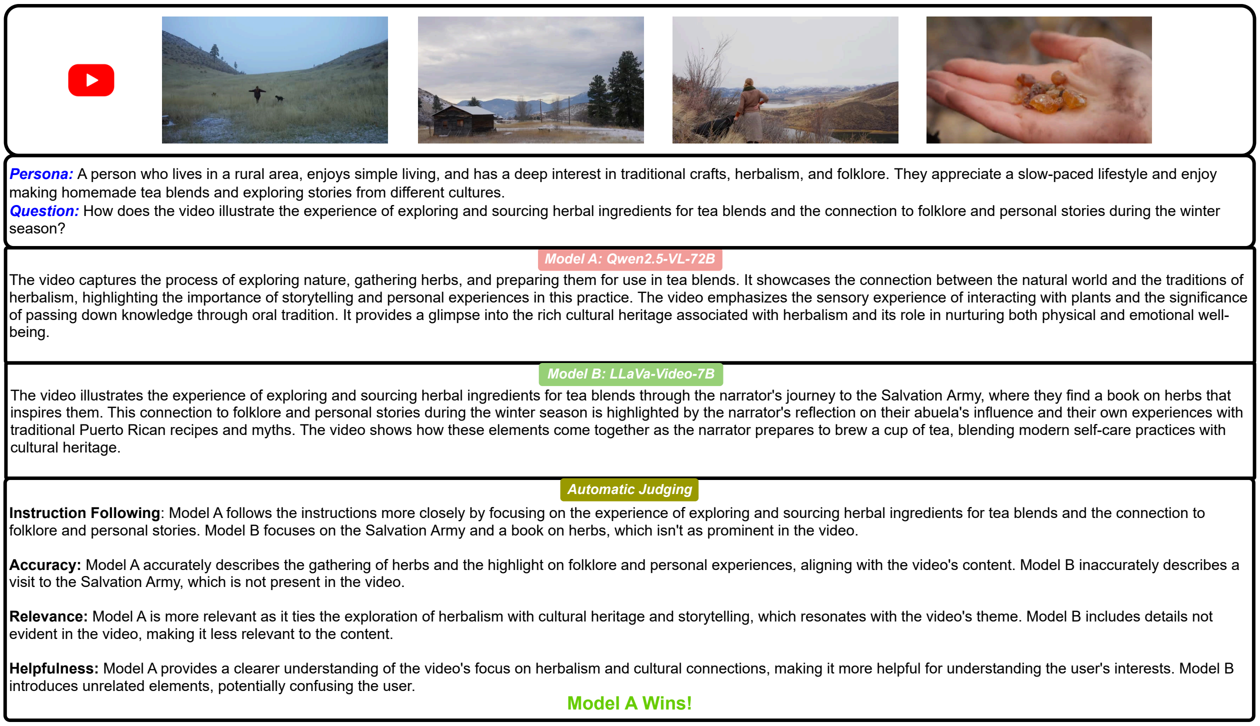
Accuracy: Model A accurately describes the steps and tools used, such as colored pencils, watercolors, and fine-tip pens. Model B accurately mentions sketching the landscape first and then adding colors but lacks detail on tools.

Relevance: Model A's response is more relevant as it provides comprehensive details that align with the user\u2019s persona as an art teacher. It gives explicit techniques and tools useful for teaching. Model B, while accurate, lacks depth.

Helpfulness: Model A is more helpful by offering detailed techniques and tool recommendations, which can be directly useful for the user's art classes. Model B is less helpful due to its brevity and lack of detail.

Model A Wins!

Figure 8. Example of the battle between GPT-4o-mini and LLaVa-Video-72B.



Persona: A person who lives in a rural area, enjoys simple living, and has a deep interest in traditional crafts, herbalism, and folklore. They appreciate a slow-paced lifestyle and enjoy making homemade tea blends and exploring stories from different cultures.

Question: How does the video illustrate the experience of exploring and sourcing herbal ingredients for tea blends and the connection to folklore and personal stories during the winter season?

Model A: Qwen2.5-VL-72B

The video captures the process of exploring nature, gathering herbs, and preparing them for use in tea blends. It showcases the connection between the natural world and the traditions of herbalism, highlighting the importance of storytelling and personal experiences in this practice. The video emphasizes the sensory experience of interacting with plants and the significance of passing down knowledge through oral tradition. It provides a glimpse into the rich cultural heritage associated with herbalism and its role in nurturing both physical and emotional well-being.

Model B: LLaVa-Video-7B

The video illustrates the experience of exploring and sourcing herbal ingredients for tea blends through the narrator's journey to the Salvation Army, where they find a book on herbs that inspires them. This connection to folklore and personal stories during the winter season is highlighted by the narrator's reflection on their abuela's influence and their own experiences with traditional Puerto Rican recipes and myths. The video shows how these elements come together as the narrator prepares to brew a cup of tea, blending modern self-care practices with cultural heritage.

Automatic Judging

Instruction Following: Model A follows the instructions more closely by focusing on the experience of exploring and sourcing herbal ingredients for tea blends and the connection to folklore and personal stories. Model B focuses on the Salvation Army and a book on herbs, which isn't as prominent in the video.

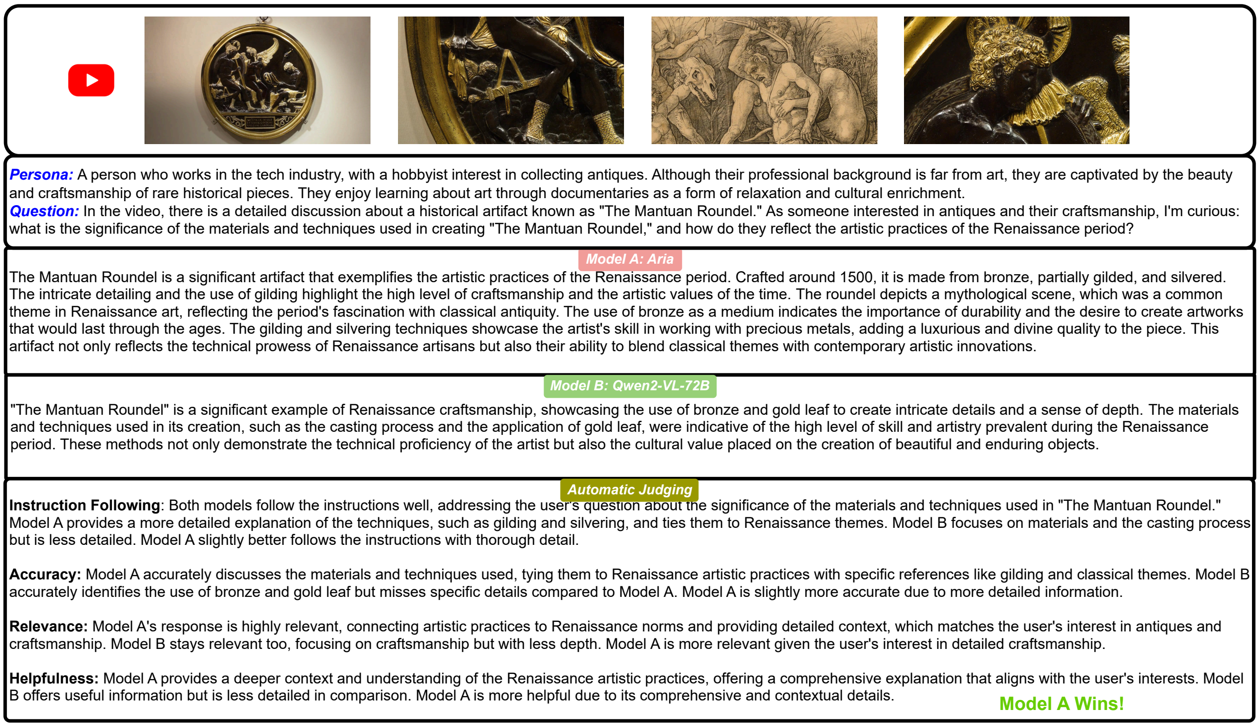
Accuracy: Model A accurately describes the gathering of herbs and the highlight on folklore and personal experiences, aligning with the video's content. Model B inaccurately describes a visit to the Salvation Army, which is not present in the video.

Relevance: Model A is more relevant as it ties the exploration of herbalism with cultural heritage and storytelling, which resonates with the video's theme. Model B includes details not evident in the video, making it less relevant to the content.

Helpfulness: Model A provides a clearer understanding of the video's focus on herbalism and cultural connections, making it more helpful for understanding the user's interests. Model B introduces unrelated elements, potentially confusing the user.

Model A Wins!

Figure 9. Example of the battle between Qwen2-VL-72B and LLaVa-Video-7B.



Persona: A person who works in the tech industry, with a hobbyist interest in collecting antiques. Although their professional background is far from art, they are captivated by the beauty and craftsmanship of rare historical pieces. They enjoy learning about art through documentaries as a form of relaxation and cultural enrichment.

Question: In the video, there is a detailed discussion about a historical artifact known as "The Mantuan Roundel." As someone interested in antiques and their craftsmanship, I'm curious: what is the significance of the materials and techniques used in creating "The Mantuan Roundel," and how do they reflect the artistic practices of the Renaissance period?

Model A: Aria

The Mantuan Roundel is a significant artifact that exemplifies the artistic practices of the Renaissance period. Crafted around 1500, it is made from bronze, partially gilded, and silvered. The intricate detailing and the use of gilding highlight the high level of craftsmanship and the artistic values of the time. The roundel depicts a mythological scene, which was a common theme in Renaissance art, reflecting the period's fascination with classical antiquity. The use of bronze as a medium indicates the importance of durability and the desire to create artworks that would last through the ages. The gilding and silvering techniques showcase the artist's skill in working with precious metals, adding a luxurious and divine quality to the piece. This artifact not only reflects the technical prowess of Renaissance artisans but also their ability to blend classical themes with contemporary artistic innovations.

Model B: Qwen2-VL-72B

"The Mantuan Roundel" is a significant example of Renaissance craftsmanship, showcasing the use of bronze and gold leaf to create intricate details and a sense of depth. The materials and techniques used in its creation, such as the casting process and the application of gold leaf, were indicative of the high level of skill and artistry prevalent during the Renaissance period. These methods not only demonstrate the technical proficiency of the artist but also the cultural value placed on the creation of beautiful and enduring objects.

Automatic Judging

Instruction Following: Both models follow the instructions well, addressing the user's question about the significance of the materials and techniques used in "The Mantuan Roundel." Model A provides a more detailed explanation of the techniques, such as gilding and silvering, and ties them to Renaissance themes. Model B focuses on materials and the casting process but is less detailed. Model A slightly better follows the instructions with thorough detail.

Accuracy: Model A accurately discusses the materials and techniques used, tying them to Renaissance artistic practices with specific references like gilding and classical themes. Model B accurately identifies the use of bronze and gold leaf but misses specific details compared to Model A. Model A is slightly more accurate due to more detailed information.

Relevance: Model A's response is highly relevant, connecting artistic practices to Renaissance norms and providing detailed context, which matches the user's interest in antiques and craftsmanship. Model B stays relevant too, focusing on craftsmanship but with less depth. Model A is more relevant given the user's interest in detailed craftsmanship.

Helpfulness: Model A provides a deeper context and understanding of the Renaissance artistic practices, offering a comprehensive explanation that aligns with the user's interests. Model B offers useful information but is less detailed in comparison. Model A is more helpful due to its comprehensive and contextual details.

Model A Wins!

Figure 10. Example of the battle between Aria and Qwen2-VL-72B.

Each sample includes a video, its video caption, and three questions. From the perspective of a user seeking AI assistance for video analysis, rank these questions.

The ranking criterion is whether the question aligns with what a real user interested in the video might ask based on their personas, rather than being a question constructed purely to test the model's capabilities. Questions ranked higher should better meet this standard.

Example
- **Video**: [Video]

- **Caption:** The video shows a man in a white shirt sitting on a sofa interacting with a doll wearing a hat. In front of them are a few glasses of beer, and several paintings are hanging on the wall. Their conversation seems to cover topics such as World War II history, Cold War information, and the development of games. The man occasionally reads from a piece of paper in his hand, while the doll appears very animated.

- **Q1:** Can you describe the sequence in which the topics are discussed in the video, including plans for the YouTube channel's future, discussions on the best tank of its era, challenges with creating early videos, and comparisons between different ships and their capabilities?

- **Q2:** As someone interested in the evolution of military equipment and its impact on modern technology, I'm curious about the main topics discussed in the video regarding military strategy and equipment development. Could you provide a summary in bullet points of the key points and any new insights about the ongoing evolution of military technology mentioned in the video?

- **Q3:** In front of the wall with the painting hanging on it, there's a chair. In front of the chair, there's a man wearing a white shirt with his left hand raised and a green toy soldier. They are looking at each other. When the subtitle mentions "something they like right yeah well then," what is the man in the shirt doing?

Ranking: Q2 > Q1 > Q3

Reasoning:
Q2 is closer to what a real user might ask about a video they are interested in. Q1 and Q3, on the other hand, are primarily designed to measure specific capabilities of the model.

Figure 11. The guideline for the question ranking annotation.

```
### Annotation Guidelines

Based on the given question and the asker's background information, compare answers A and B across four dimensions and overall: **A, B, Tie Good, Tie Bad**. The four dimensions are **Instruction Following, Accuracy, Relevance, and Helpfulness**. Refer to the following criteria for evaluation:

### Evaluation Dimension Criteria

#### **1. Instruction Following**
- **Description:** Ensure the response adheres to the user's instructions.
- **Criteria:**
  - Fails to meet the core expectations in the user's query or instruction.
  - Errors in context understanding lead to incorrect output. For ambiguous instructions, there is no correct answer in the response.
  - Inaccurate content or factual errors, including hallucinations. Misuse of technical terms or ambiguous expressions.
  - Irrelevant content that does not address the user's query.
  - Incorrect or invalid references in the response.

#### **2. Accuracy**
- **Description:** Accurately utilizes information from the video to answer the question without fabrication or misrepresentation.
- **Criteria:**
  - Generated content must be factually correct and free of hallucinations.
  - The response should be contextually consistent, with accurate and unambiguous use of technical terms and knowledge.

#### **3. Relevance**
- **Description:** Considers the asker's background information, ensuring that the output aligns closely with the user's needs and directly addresses the task or question.
- **Criteria:**
  - Avoids irrelevant or off-topic answers.
  - Responses are complete, detailed, and in-depth, avoiding overly simplistic or single-perspective answers.

#### **4. Helpfulness**
- **Description:** The response should effectively meet the user's needs and provide valuable information.
- **Criteria:**
  - Helps the user better understand or resolve their issue.
  - Avoids providing irrelevant or vague answers.
```

Figure 12. The guideline for the response judging annotation.