CineVerse: Consistent Keyframe Synthesis for Cinematic Scene Composition Supplementary Material

Quynh Phung[1*] Long Mai[2] Fabian David Caba Heilbron[2]
Feng Liu[2] Jia-Bin Huang[1] Cusuh Ham[2]

[1] University of Maryland, College Park [2] Adobe Research
{quynhpt,jbhuang}@umd.edu {malong, caba, fengl, ham}@adobe.com

1. Additional qualitative results

We include additional qualitative comparisons of our method against the baselines shown in Fig. 3 and Fig. 4. Our approach generates coherent keyframes with superior prompt alignment, consistent character portrayal, and precise shot sizing. It continues to perform robustly even as the number of shots increases, preserving both character consistency and overall narrative coherence.

Fig. 5 illustrates some of the issues with IC-LoRA [7], such as generating cropped images or misalignment with the input prompt. These problems are due to their data processing pipeline.

2. Ablation studies

To evaluate the importance of individual attributes in shot-level annotations, we conduct an ablation study by systematically removing one attribute at a time during generation (in Fig. 1). Our findings highlight that:

- Without shot descriptions, the model relies solely on scene-level text and generates visually similar shots, often failing to convey distinct actions or meaningful scene progression.
- Omitting character descriptions leads to inconsistencies in character appearance and identity across shots, reducing coherence in multi-shot scenes.
- Excluding shot size impacts the intended visual emphasis. For example, actions such as a character jumping lose their dramatic effect without a wide or long shot that captures the full scope of motion and surrounding context.

These results demonstrate that each attribute contributes to aligning the generation with cinematic composition principles, and removing any one of them degrades the fidelity and clarity of the generated sequences.

Our ablation studies systematically investigated the impact of four key hyperparameters on model performance, as shown in Table 1: LoRA rank, training iteration count, addition of borders between training shots, and scene/shot balancing. Among LoRA configurations, a rank of 128 yielded the highest accuracy (88.83%) and best CLIP score (0.2118), indicating optimal capacity for feature adaptation. Analyzing the number of training iterations, we observed that performance improved steadily as iterations increased up to 16k steps, achieving peak accuracy (88.83%), but notably declined at 20k steps, suggesting potential overfitting. Including explicit borders between training shots significantly enhanced accuracy (88.83% vs. 47.20%) and improved CLIP scores (0.2121 vs. 0.1960), highlighting the importance of clear visual delineation. Lastly, training on balanced scene/shot data further boosted accuracy to 88.83%, emphasizing that balanced training data is crucial for accurate shot-count generation. Consequently, the optimal hyperparameter combination identified was LoRA rank 128, training for 16k iterations, inclusion of borders between shots, and using balanced data.

3. Dataset discussion

We depict the structure of movies in Fig. 6.

To leverage LLaVA-OneVision-72B [22] for our visual understanding task, we provide a straightforward instruction format followed by images. There are three main extracted information: setting, shot description, and character description. An example instruction prompt for generating the setting, character descriptions, and per-shot descriptions is shown in Fig. 8. Given the input frames and detailed instructions, LLaVA-OneVision demonstrates strong task comprehension. As a result, our new dataset, constructed using MLLMs, provides richer shot-level contextual information than existing datasets, as shown in Table 2.

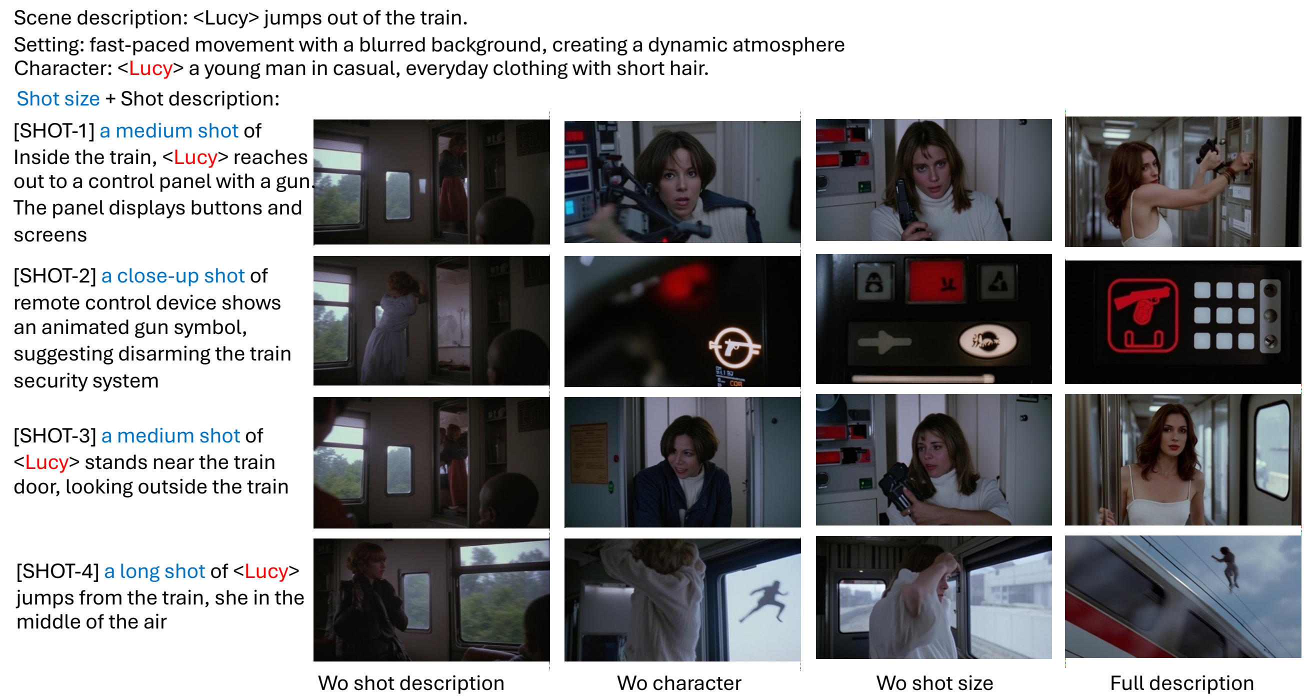


Figure 1. **Ablation study on scene attributes**. We assess the impact of removing individual shot-level attributes during generation. Without shot descriptions, generated frames are nearly identical and fail to capture distinct actions. Removing character descriptions reduces consistency across shots, while omitting shot size diminishes visual intent—for example, long shots are crucial to convey the full context of dynamic actions like jumping.

Method	CLIP ↑	DS ↓
Ours	0.285	0.545
1P1S [17]	**0.294**	**0.323**

Figure 2. **Analyze the automatic metrics.** Our method produces scenes with stronger nar- rative alignment and character consistency, even when the metrics favor 1P1S [17].

4. Additional evaluation metrics and limitation of current automatic metrics

Cinematic scene composition is a novel task without reliable automatic metrics. Existing metrics for generative tasks have known limitations, such as failing to capture fine-grained consistency in character identity and setting, key factors for coherent storytelling. Additionally, these metrics often favor minor visual changes across frames, whereas diversity in shot sizes and perspective is an essential component for compelling visual storytelling. In the example in the Fig. 2, our method produces scenes with stronger narrative alignment and character consistency, even when the metrics favor 1P1S [17]. This highlights the need for better human-aligned evaluations. We include automated metrics in our evaluations for completeness.

Table 1. Ablation studies on different training settings.

	Acc.(%) ↑	CLIP ↑	DS↓
LoRA rank/alpha			
32	73.10	0.2048	**0.4145**
64	77.83	0.2073	0.4539
128	**88.83**	**0.2118**	0.4476
Training iteration			
2k	75.24	0.2085	0.3853
5k	77.15	0.2068	0.3998
10k	79.03	**0.2165**	0.5501
15k	87.58	0.2099	0.4499
16k	**88.83**	0.2118	0.4476
20k	72.42	0.2018	**0.3476**
Adding border between training shots			
×	47.20	0.1960	**0.4322**
	88.83	**0.2121**	0.4476
Scene/shot balancing			
×	57.86	0.1993	0.47612
	88.83	**0.2118**	**0.4476**

MLLM evaluations. We leverage LLaVA-OneVision

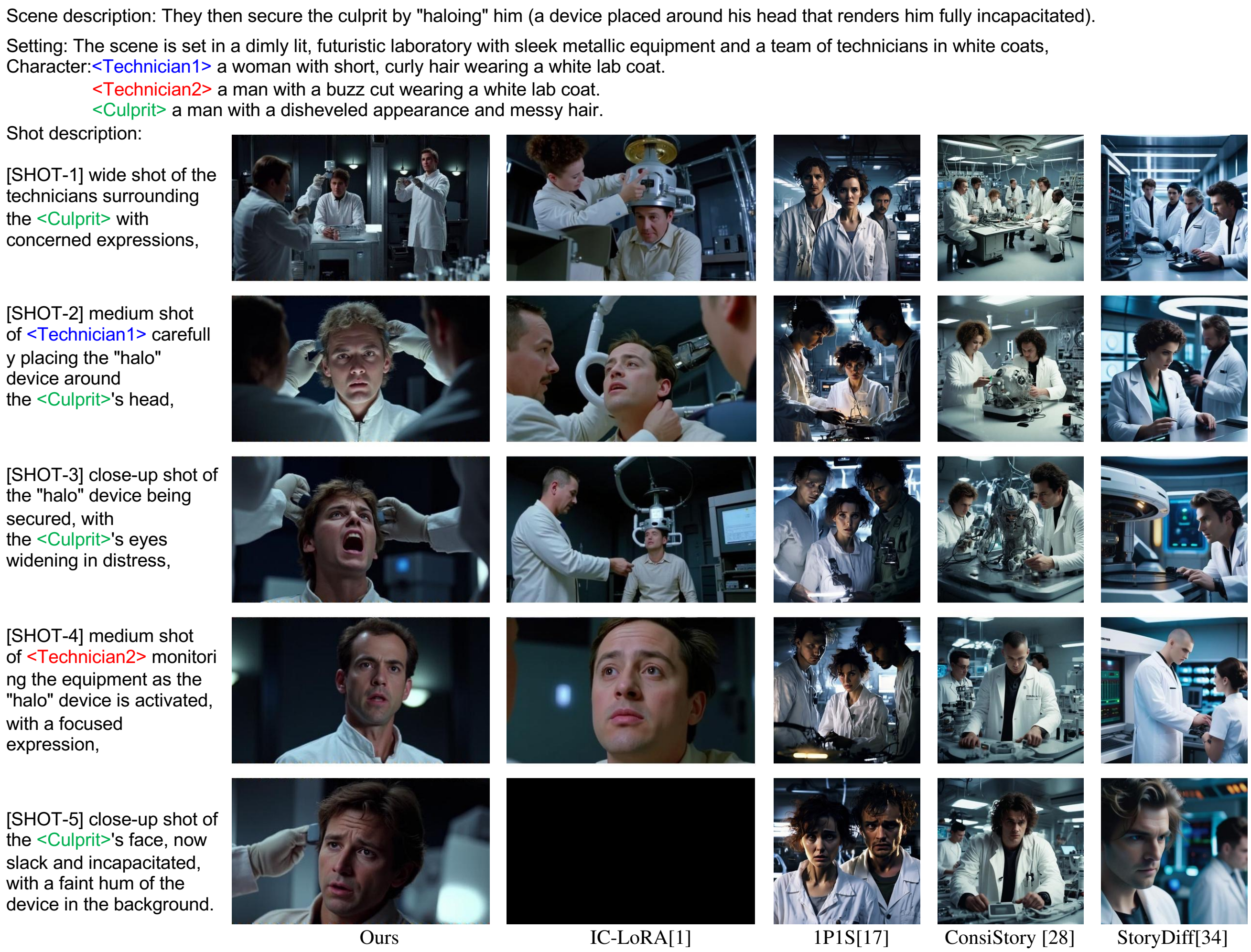


Figure 3. **Additional visual comparisons with state-of-the-art multi-shot image generation.** Our approach generates coherent keyframes with superior prompt alignment, consistent characters, and precise shot sizing.

Table 2. Comparison of CineVerse with other datasets: CondensedMovie, MSA, MovieNet, Storyboard20K. Our dataset is rich with attributes at the shot level.

Dataset	# movie	# scene	# shot	Shot Desc.	Char. Desc.	Setting	Cam. Shot
CondensedMovie	3.6K	33K	400K	×	×	×	×
MSA	327	4.5K	-	×	×	×	×
MovieNet	1.1K	43K	3.9M	×	×	×	×
Storyboard20K	400	20K	150K	×	×	×	×
CineVerse	312	10K	46K				

[22], a state-of-the-art MLLM capable of understanding image sequences through visual narrative analysis, for evaluations. Additionally, prior work has demonstrated that GPT-4o can evaluate image sequences with strong alignment to human judgment. Thus, we use both GPT-4o and LLaVA-OneVision to assess the performance of our method for the cinematic scene composition task. As shown in Table 3, our method consistently outperforms all baselines across evaluation metrics, with similar trends observed in both GPT-4o and human evaluations.

Since MLLMs can assess multiple aspects of a visual sequence, we expand our evaluation beyond the initial four criteria to include two additional ones:

1. **Action flow**: Assess whether the sequence displays a smooth and logical progression of actions and expressions that reflect the scene's dynamics.

2. **Camera movement**: Determine whether transitions between keyframes resemble coherent, movie-like camera motions that enhance storytelling.

We evaluate 200 images per baseline. Following the structure of the user study, each question compares our method against a single baseline. To mitigate biases and ensure fairness, we provide clear instructions to guide the MLLM in

Scene description: When schools start to decline his applications, Toby realizes he may have to stay in Concrete like his friends and gets a job in a grocery store.
Setting: The scene is set in a small, rundown town with a declining economy, featuring a local grocery store with faded signs and worn-out streets.
Character: <Toby> a young man in casual, everyday clothing with short hair.
<Store Customers> individuals of various ages and genders in everyday attire.
<Store Owner> a middle-aged man in a simple, store-owner uniform with a receding hairstyle.
Shot description:

[SHOT-1] wide shot of the small main street, showing the grocery store and other rundown buildings,

[SHOT-2] medium shot of <Toby> walking down the main street, looking concerned and unsure about his future,

[SHOT-3] close-up shot of <Toby> holding a stack of declined job applications, looking disappointed and frustrated,

[SHOT-4] wide shot of the grocery store's interior, with <Toby> standing at the counter, talking to the <Store Owner>

[SHOT-5] medium shot of <Toby> putting on a grocery store uniform, indicating he has taken the job,

[SHOT-6] close-up shot of <Toby> restocking shelves, looking determined to make the best of his situation,

[SHOT-7] wide shot of <Toby> interacting with <Store Customers>, smiling and assisting them, as he settles into his new role.

Figure 4. **Additional visual comparisons with state-of-the-art multi-shot image generation**. Our approach performs well even with a larger number of shots, maintaining character consistency and narrative coherence.

selecting the better image sequence for each criterion and randomize the side on which each method appears for all trials. The full instruction prompt is shown in Fig. 11.

User study. Participants first receive detailed instructions at the beginning of the survey, including task explanations and illustrative examples distinguishing good from bad cases (an example can be found Fig. 9). These initial instructions ensure better comprehension of the survey aspects. Additionally, each question in the survey has a time limit, helping to reduce noise by preventing users from lingering excessively on difficult-to-decide questions.

Additional metrics. We evaluate the accuracy of generating the correct number of images in the scene, as specified by the input scene plan. A scene is considered "correct" if its shot count matches the expected number. For sequences generated by IC-LoRA, where images lack clear borders between shots, we estimate transitions between frames by calculating the pixel difference between adjacent rows, with

Figure 5. **IC-LoRA common failures**. IC-LoRA often generates cropped frames and/or the incorrect number of images.

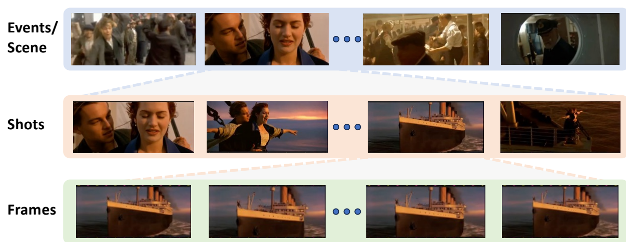


Figure 6. **Movie structure.** A movie is composed of unique scenes and events that drive the storyline. Each scene consists of multiple shots establishing context, highlighting character emotions, or emphasizing key details. At the finest level, individual frames bring these shots to life.

Figure 7. **Limitations.** Our method sometimes still suffers from bad image quality with artifacts, such as missing borders, and mismatches with the shot size specified in the shot description.

the highest difference indicating the separation boundary. Since we incorporate distinct borders to separate the frames for training CineVerse, we utilize the Canny edge detector to identify these borders.

Our method shows significant improvements in generating the correct number of shots compared to IC LoRA [**?**](see Table 4). For instance, ours achieves 95.45% accuracy compared to 34.84% for three frames and similarly across higher frame counts, demonstrating robustness.

Table 3. Comparison of our method and the baselines using LLaVA-OneVision.

	Textual Align.		Consistency		Continuity	
Ours vs.	Scene	Shot	Char	BG	Action	Camera
1P1S	82.34	82.32	81.43	82.34	83.52	82.43
ConsiStory	65.45	63.63	65.45	63.63	65.45	65.45
StoryDiff	82.43	83.42	82.43	84.56	83.45	83.45
IC-LoRA	73.23	71.14	74.22	70.21	69.53	74.23

Table 4. Comparison of shot generation accuracy.

# shots	3	4	5	6	7	8	9	10
IC LoRA	34.84	37.60	17.64	16.67	26.08	21.05	08.33	16.66
Ours	95.45	91.45	91.17	85.98	78.87	72.72	60.24	42.24

5. Prompting LLMs for scene planning

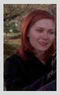
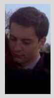
Fig. 10 shows the full instructions used in the planning stage, guiding the LLM to produce outputs with the correct format and accurate attributes for generation. Fig. 11 presents the complete evaluation prompts used by GPT-4o and LLaVA-OneVision [22] to assess the consistency and quality of generated image sequences within the same scene.

6. Limitation and discussion

We have improved textual alignment and image separation; however, our method still suffers from artifacts, missing borders, and occasional mismatches with the text prompt, as depicted in Fig. 7. We plan to address these issues in future work.

Shot description generation

The characters in the scene:

Mary Jane:  , Peter 

Here is list of images in a movie scene and corresponding character names in each image:

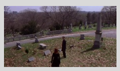
Image 1: Mary Jane, Peter, 

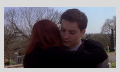
Image 2: Mary Jane, Peter 

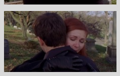
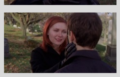
Image 3: Mary Jane, Peter 

Image 4: Mary Jane, Peter 

In the context of the story: "Mary Jane finds Peter Parker at *Uncle Ben's grave* and confesses her love for him ". Describe each image in the context of story , using the given character names and their portraits, focusing more on the actions of given character and the setting. If frame does not have characters, no need to mention character names.' Follow the structure below for formatting:

Image descriptions:

Image 1: [Image description]

Image 2: [Image description]

Output

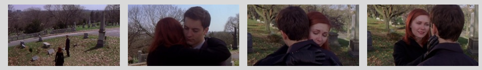
Image 1: Mary Jane and Peter stand in a cemetery, with Mary Jane embracing Peter from behind.

Image 2: Mary Jane and Peter are in a close embrace, surrounded by tombstones and bare trees. Peter looks down at Mary Jane with a serious expression.

Image 3: Mary Jane and Peter hug each other tightly in the cemetery, with Mary Jane crying and Peter holding her close.

Image 4: Mary Jane and Peter stand face to face, with Mary Jane smiling and Peter looking concerned, surrounded by tombstones and trees.

Setting generation

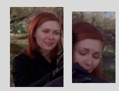


From sequence of images in one clip, describe the place where they capture sequence of image, focusing on describing the background set up of scene, don't focus on people, focusing on important things, and providing the general overview, within 35 words. Start with: The scene depicts ...

Output

The scene depicts a serene cemetery scene with a calm atmosphere, set in late autumn or early spring with clear skies.

Character description generation

Here are images of a character: 

The above images depict one person, please briefly describe the character including: appearance, gender, hair style and color, and clothes. Start with: The character is

Output

The character is a young woman with long, flowing red hair, wearing a black dress.

Refining the scene description

Please refine the captions in the context of the story such that, when a refined caption stands alone, people can still understand the event.

Story: "Peter goes to visit Uncle Ben's grave. Mary Jane finds him there and confesses her love for him. She kisses him tenderly, passionately. Peter wants to tell her the truth but can't. Instead, he tells her he can never be more than her friend. Mary Jane has an inkling that she might have kissed him before but Peter walks away, knowing both his blessing and his curse in life: "Who am I? I'm Spider-Man."

The caption: Mary Jane finds him *there* and confesses her love for him.

Note that you should return one refined caption, which should specify the character's name and remain clear when removed from the story, minimizing changes compared to the original caption.

Also indicate which words were replaced by character names (e.g., "he" replaced by "Lucy"). Return the result as a list of corresponding word pairs, for example: [[he, Lucy], ...]

Output

Mary Jane finds Peter Parker at *Uncle Ben's grave* and confesses her love for him. [[him, Peter Parker]]

Figure 8. **Instruction to extract data attributes using LLaVa-OneVision**. We provide detailed instruction for 4 tasks: shot description, setting, character description generation and refining scene description.

Task: Given 2 images, we need your help to select which better depicts the input text

How to evaluate:

The image should clearly depicts the text description, including characters' action, expression, shot size, ...
We provide a good example and a bad example of shot alignment here:

Good example	Bad example
 **Shot description: close-up shot of Lily's face, highlighting her enthusiastic and hopeful expression.** *This image clearly depicts the described character and emotion, as well as the shot size (close-up)*	 **Shot description: close-up shot of Globin's hands as he releases a barrage of pumpkin bombs.** ***This is bad because the image is a medium shot rather than a close-up, and key information outlined in the text is missing e.g., Globin's hands, pumpkin bombs.***

Figure 9. **User study**. The instructions presented to users at the beginning of the survey.

Objective: As a movie director, your task is to carefully plan shots that effectively communicate the scene's narrative visually. Consider the following guidelines to ensure precision, clarity, and cinematic professionalism in your shot planning:
Output Format:
1.Your final output must strictly adhere to the following structure:
[MOVIE-SHOTS] The scene is set in [background description]. <Character1> character details, <Character2> character details, ...,
[SCENE-1] description of shot 1,
[SCENE-2] description of shot 2,
...
2.Character1, Character2, etc.: the characters appear in the scene, the name should in side <>.
3. character details:
Include only the following details for each character (if possible):
Outfit
Gender
Hairstyle
Do not add any further character details.
4. description of shot 1, description of shot 2, etc: the detailed shot planning for each shot, based solely on the scene description. indicate:
The shot number (e.g., [SCENE-1], [SCENE-2], etc.)
The type of shot (Only including: wide shot, medium shot, close-up shot)
A description of the character's action, expression or focus as directly described in the scene description, not describe the dialog in shot description.
5. Example Input:
Scene description: Wladyslaw later blends in with the ten percent or so of the Jews that the Nazis kept alive in the ghetto to use for slave labor, tearing down the brick walls separating the ghetto and rebuilding apartment houses for new, non-Jewish residents.
Plan 3 shots for the above scene:
6. Example Output:
[MOVIE-SHOTS] The scene is set in a desolate, war-torn urban environment with partially destroyed brick walls and emerging construction sites under a gloomy sky. <Wladyslaw> a man in worn, laborer's clothing with unkempt hair, <Jewish Laborers> individuals in similar worn clothing,
[SCENE-1] wide shot of <Wladyslaw> blending in with a group of <Jewish Laborers> tearing down brick walls,
[SCENE-2] medium shot focusing on <Wladyslaw> as he actively participates in the demolition,
[SCENE-3] close-up shot of a <Jewish Laborer> diligently rebuilding an apartment facade

Figure 10. **Scene planning instruction prompt**. Example of a prompt used to guide LLMs in the *scene planning* stage of CineVerse.

```
You are an expert in movie scene analysis. You will be given 2 sequence of images,
    the two representing one scene from a movie. Your job is to evaluate and
    select the better sequence that best exemplifies the scene based on three
    specific criteria.
    1. Textual Alignment:
        * Overall Scene: Assess how well the keyframes capture the narrative, mood
          , and setting as described in the overall scene description.
        * Shot Details: Evaluate how accurately each keyframe reflects the
          detailed descriptions provided for individual shots.
        * Key Points: Consider whether the depicted actions, expressions, and
          visual details align with both the story and shot specifics
    2. Consistency:
        * Character Consistency: Ensure that the main character's appearance (
          clothing, hairstyle, facial features) remains uniform across all
          keyframes, even as their actions vary.
        * Background Consistency: Verify that the backgrounds, although possibly
          shown from different perspectives, clearly indicate the same location.
    3. Continuity:
        * Action Flow: Analyze the sequence for smooth and logical progression of
          actions and expressions that mirror the described scene's dynamics.
        * Camera Movement: Evaluate if the camera transitions and shifts between
          keyframes create a coherent, movie-like progression that enhances the
          storytelling
    For each answer, you should explain why you choose this option.
    Then the final answer should be the choosen sequence (the best sequence) for
       each aspect like bellow format, the choosen sequence can be different for
       different aspects:
    1. Textual Alignment:
        * Overall Scene: [chosen sequence]
        * Shot Details:[chosen sequence]
        * Key Points: [chosen sequence]
    2. Consistency:
        * Character Consistency: [chosen sequence]
        * Background Consistency: [chosen sequence]
    3. Continuity:
        * Action Flow: [chosen sequence]
        * Camera Movement: [chosen sequence]
```

Figure 11. **Evaluation instruction prompt.** Instructions for GPT-4o and LLaVA-OneVision to assess the results of the keyframe generation stage.