CineVerse: Consistent Keyframe Synthesis for Cinematic Scene Composition Supplementary Material

Anonymous ICCV submission

Paper ID 12782

Figure 7. **Limitation.** Our method sometimes still suffers from bad image quality with artifacts, such as missing borders, mismatching with camera shot in shot description. .

Figure 8. **Visual comparisons with IC-LoRA** . IC-LoRA often generates cropped frames and/or the incorrect number of images, and fails to respect the specified camera shot size.

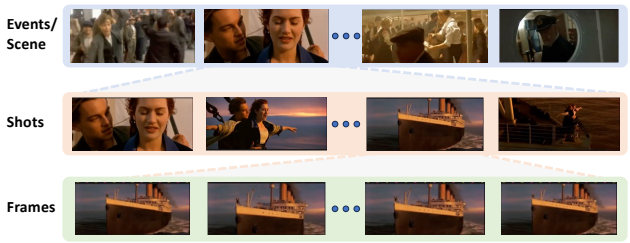

Figure 9. **Movie structure.** A movie is composed of unique scenes and events that drive the storyline. Each scene consists of multiple shots establishing context, highlighting character emotions, or emphasizing key details. At the finest level, individual frames bring these shots to life. Our work aims to empower everyday users to composite cinematic scenes at the shot level.

7. More qualitative results

We provide more qualitative results to compare ours and baselines in the file visualization.html

Fig. 8 show some problem of In-Context LoRa paper, like generating cropped image or mis alignment with input prompt. which mention in main paper.

8. 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. We plan to address these issues in future work.

9. Ablation studies

Our experiments indicate that fine-tuning for approximately 16k steps best balances task adaptation (refer to Table 5).

This modification improves the accuracy of shot count generation (see the third part in Table 5). (We can see the vary of height can affect the model in the Table 5 in size category).

Ablation studies show that balancing shot numbers is crucial—models trained on unbalanced data generate inaccurate shot counts, as demonstrated in Fig. 5.

10. Dataset discussion

Prompting using for LLaVa-Onevision : to ask llava one vision to do the task uderstanding task, we just simple instruction LLava-Onevision in the struction like that format : scene description plus **In the context of the given story: . Describe each image in the context of story , using the given character names and their portraits, fo-**

Table 5. Ablation studies on different training settings.

	Acc.(%) ↑	CLIP ↑	DS↓
LoRA rank/ alpha			
32	73.10	0.2048	**0.5855**
64	77.83	0.2073	0.5461
128	**88.83**	**0.2118**	0.5524
Training iteration			
2k	75.24	0.2085	0.6147
5k	77.15	0.2068	0.6002
10k	79.03	**0.2165**	0.5501
15k	87.58	0.2099	0.5641
16k	**88.83**	0.2118	0.5524
20k	72.42	0.2018	**0.6524**
Adding border between training shots			
×	47.20	0.1960	**0.5678**
✓	**88.83**	**0.2121**	0.5524
Scene/shot balancing			
×	57.86	0.1993	0.5238
✓	**88.83**	**0.2118**	**0.5524**

Table 6. Accuracy between our text-to-image and IC-LoRA. with different number of shots.

# 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

Table 7. Compare CineVerse datset and others datasets: condensedMovie, MSA, MovieNet, Storyboard20k. Our dataset with rich attributes in 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	✓	✓	✓	✓

cusing more on the actions of given character and the setting. If frame does not have characters, no need to mention character names" THen ir can understand the task well because we use llava-Onevision 72b." We compare serval datasets and see that our proposed dataset contain rich information in shot-level compare other Table 7

11. Additional evaluation 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 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.

We show the comparasion of our method and In-Context LoRa in Table 6

MLLM evaluation : We find LLaVa-Onevision, the state-of-the-art MLLM model, that understand sequences of images—is their ability to analyze visual narratives. Previous work has shown that GPT-4 can evaluate image sequences with results that align well with human judgment. Building on this, we employ GPT-4 along with llava onevision to assess the outcomes of our multi-shot keyframe generation method. We compare ours and baselines in Table 8, anf see our method outperform other baslines in all aspects, showing similar trend in GPT-4 and human evaluation.

Because LVMs can evaluate multiple aspects of a visual sequence, we extend our analysis beyond the four initial aspects to include two additional criteria:

1. **Action Flow** Analyze the sequence for smooth and logical progression of actions and expressions that mirror the described scene's dynamics.
2. **Camera Movement**: Evaluate if the camera transitions and shifts between keyframes create a coherent, movie-like progression that enhances the storytelling

We use 200 images for each baseline comparision, Similar to user study, we compre our method and one of baseline in each question. We use the thoughtful instruction to guide LVMs how to choose the best sequence of image base on each aspect. We also applied some technique like random the order of image to avoid bias in oder. The instruction can be find in the supplement material. We evaluate ours and baseline in main sup also conpare

Table 8. Comparison of ours method and 5 baslines 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

MLLM evaluation : Additionally, we leverage GPT-4 and LLaVa-OneVision, a state-of-the-art MLLM model renowned for its ability to understand image sequences and analyze visual narratives. Prior work has demonstrated that GPT-4 can evaluate image sequences in a manner that

closely aligns with human judgment. Building on these findings, we employ GPT-4 in combination with LLaVa-OneVision to assess the performance of our approach for the cinematic scene composition task.

In terms of user protocol, user first will se the instruction in the beginning of surevey., that explain the task and a good and bad example of the aspect in survey, this may help user understand more about the task. The during the doing survey, user have a limitation time for each question , this may help to reduce the noise in survey when user ignore question is hard to decide.

12. Prompts using for LLMs planning and GPT-4 evaluation

The Fig. 10 provides the full LLM instruction at the shot planning stage. It helps the model generate right format and correct attribute for generation stage. The Fig. 11 provide full prompt for GPT-4 and LLaVa-Onvision to evaluate the sequence of images in the same scene.

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. **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]
    Note that: evaluate each aspect independently, and favor the sequence look
       like realistic movie keyframe (not cartoon or digitial style) and have
       border separate images.
```

Figure 11. Instruction for GPT-4V and llaVa-Onevision to asses the visual instruction