

From Sound to Sight: Towards AI-authored Music Videos

Supplementary Material

6. Prompt Flow

6.1. CLAP-based Approach

In the following, we present the prompt structure for the CLAP-based approach in pseudocode in Block 1, an example prompt in Block 2, and the corresponding response in Block 3.

Block 1. Prompt structure for the LLM used in the CLAP-based approach

```
number_of_scenes = length(segments)

final_prompt = "You need to think of a story and a script for a music video. {additional_prompt}
The story structure needs to have a beginning, middle and an ending.
You will write the story based on the characteristics of the music that are provided to
you.
The story needs to be told as descriptions of scenes that will appear in a music video
.
The structure needs to be reflected in the scene layout, to try to mimic the story's
progression.
The story should include at least one animal character. For consistency, repeat the
animal's description in every scene.
Make sure the scene descriptions are concrete and to the point, they need to be easy
for a text2video model to generate videos from.
There are {number_of_scenes} scenes in total.
When listening to the entire song, it can be described like this:
the overall {overall_song_description_keywords}
When listening to individual segments, they can be described like this:
{scene_descriptions}
Each scene needs to visually match the following guidelines:
{overall_song_description_keywords_visual_guides}
Scene descriptions need to be as brief as possible. Every scene can be described in one
sentence at the most. Omit all unnecessary details from the description.
In your response, start the description of every scene with the exact letters: "SCENE
#:", '#' substituted with the scene number. Do NOT add any special or any other
type of characters to this line! Example: "SCENE 1:\n"
Before the start of the script, add the words "BEGIN SCRIPT\n" so that I can easily
extract it."
```

Block 2. Prompt sent to the LLM based on CLAP features

You need to think of a story and a script for a music video.
The story structure needs to have a beginning, middle and an ending.
You will write the story based on the characteristics of the music that are provided to you.
The story needs to be told as descriptions of scenes that will appear in a music video.
The structure needs to be reflected in the scene layout, to try to mimic the story's progression.
The story should include at least four people. For consistency, repeat their descriptions in every scene.
Make sure the scene descriptions are concrete and to the point, they need to be easy for a text2video model to generate videos from.
There are 7 scenes in total.
When listening to the entire song, it can be described like this:
the overall Instrumental energy is It has multiple peaks and valleys throughout.. Instrumental palette is Orchestral or cinematic instruments. Tempo range is Very fast (140+ BPM). Production quality is Very polished, glossy, and modern. Mood is Uplifting and bright.

When listening to individual segments, they can be described like this:

Scene 1:
Instrumental intensity is Moderate intensity with a clear beat. Prominent element is String or orchestral elements. Dynamic shift is It fluctuates multiple times within the segment. Rhythm is Irregular or changing time signatures. Function in transitioning is It acts as a noticeable break or "breather". The scene will be 5.49 seconds long.

Scene 2:
Instrumental intensity is Moderate intensity with a clear beat. Prominent element is String or orchestral elements. Dynamic shift is It fluctuates multiple times within the segment. Rhythm is Irregular or changing time signatures. Function in transitioning is It features a sudden drop or pause before the next section. The scene will be 7.13 seconds long.

Scene 3:
Instrumental intensity is High energy and full instrumentation. Prominent element is String or orchestral elements. Dynamic shift is It fluctuates multiple times within the segment. Rhythm is Irregular or changing time signatures. Function in transitioning is It features a sudden drop or pause before the next section. The scene will be 7.87 seconds long.

Scene 4:
Instrumental intensity is High energy and full instrumentation. Prominent element is String or orchestral elements. Dynamic shift is It fluctuates multiple times within the segment. Rhythm is Irregular or changing time signatures. Function in transitioning is It cleanly continues the energy from the previous segment. The scene will be 6.66 seconds long.

Scene 5:
Instrumental intensity is Moderate intensity with a clear beat. Prominent element is String or orchestral elements. Dynamic shift is It stays uniformly loud/energetic. Rhythm is Irregular or changing time signatures. Function in transitioning is It slowly fades out or prepares for a drop. The scene will be 6.2 seconds long.

Scene 6:
Instrumental intensity is High energy and full instrumentation. Prominent element is Synths or electronic sounds. Dynamic shift is It fluctuates multiple times within the segment. Rhythm is Irregular or changing time signatures. Function in transitioning is It dramatically shifts the energy or mood. The scene will be 5.72 seconds long.

Scene 7:
Instrumental intensity is High energy and full instrumentation. Prominent element is String or orchestral elements. Dynamic shift is It fluctuates multiple times within the segment. Rhythm is Irregular or changing time signatures. Function in transitioning is It features a sudden drop or pause before the next section. The scene will be 4.94 seconds long.

Each scene needs to visually match the following guidelines:
Location is exterior. Visual setting is Natural. Visual style is Monochromatic or limited color palette . Visual focus is Multiple focal points or characters.

Scene descriptions need to be as brief as possible. Every scene can be described in one sentence at the most. Omit all unnecessary details from the description.
In your response, start the description of every scene with the exact letters: "SCENE #:", '#' substituted with the scene number. Do NOT add any special or any other type of characters to this line! Example: "SCENE 1:\n"
Before the start of the script, add the words "BEGIN SCRIPT\n" so that I can easily extract it.

Block 3. Prompt response from LLM based on CLAP features

Okay, so I'm trying to help this user who wants a story and a script for a music video. Let me break down what they need. They provided a lot of specific details about the song's structure, the scenes, and the visual guidelines.

First, the story needs to have a beginning, middle, and ending. It should be told through scene descriptions, and each scene needs to reflect the song's segments. They mentioned that there are seven scenes, each with specific characteristics like instrumental intensity, prominent elements, dynamic shifts, and transition functions. Plus, each scene should be a single sentence, starting with "SCENE #:".

The user also provided guidelines for the visual style: exterior locations, natural settings, monochromatic or limited color palette, and multiple focal points or characters. The story should include at least four people, with consistent descriptions across scenes. The production quality is polished and modern, with an uplifting and bright mood.

I need to make sure each scene's description is concrete and easy for a text-to-video model to generate. The scenes should transition smoothly, reflecting the dynamic shifts and energy changes of the song. Also, the locations should be natural, like forests, mountains, or beaches, to fit the exterior requirement.

I'll start by outlining each scene based on the provided structure, ensuring that each one matches the corresponding segment's characteristics. I'll include four main characters, each with distinct roles and appearances, to keep the story consistent. The descriptions should be brief, focusing on the main visual elements without unnecessary details.

Finally, I'll structure the script with "BEGIN SCRIPT" at the start and each scene clearly labeled. I'll make sure the transitions between scenes follow the energy and mood shifts indicated by the song's segments. This should result in a cohesive and visually engaging music video script that aligns perfectly with the provided song details.

</think>

BEGIN SCRIPT

SCENE 1: A group of four people, dressed in vibrant, flowing outfits, stand at the edge of a vast, sun-drenched forest. Strings of light shimmer in the distance as they begin to walk forward, their movements synchronized to the rhythm.

SCENE 2: The group reaches a clearing where a lone figure, dressed in a shimmering silver cloak, stands. The figure raises their arms, and the strings of light twist into a swirling vortex above. The group steps into the light, their colors blending together as the strings pull them toward the sky.

SCENE 3: The group emerges from the light into a futuristic, neon-lit cityscape. Skyscrapers glow with shifting colors, and a massive screen displays abstract patterns. The lone figure from before now stands at the center, their cloak glowing faintly. The group surrounds them, their movements now more chaotic and energetic.

SCENE 4: The city transforms around them. Buildings collapse into sand, and the sky darkens slightly. The lone figure raises their hands, and the sand reforms into a beach with crashing waves. The group stands amidst the waves, their expressions reflective and determined.

SCENE 5: A single figure, now isolated on the beach, begins to fade into the horizon. The group watches, their forms becoming more abstract as the scene transitions. The last shot lingers on the group, their forms dissolving into light as the music fades.

SCENE 6: The group reappears in a serene mountain valley. Snow-covered peaks rise into the distance, and a soft glow surrounds them. The lone figure, now wearing a different outfit, leads the group toward a crystal cave at the base of the mountains.

SCENE 7: The group enters the cave, the walls shimmering with faint, ethereal light. The lone figure raises their hands, and the cave fills with a blinding light. The group steps forward, their forms dissolving into light as the music crescendos to a final, uplifting note.

END SCRIPT

6.2. LALM-based Approach

Block 4 presents a prompt used when inquiring a LALM model to write a video script. Block 5 is an example of a final video script generated by the LALM.

Block 4. Analysis prompt for the LALM model

```
lalm_conversation = [
    {'role': 'system', 'content': 'You are a helpful assistant.'},
    {"role": "user", "content": [
        {"type": "audio", "audio_url": song_audio},
        {"type": "text", "text": f"Think of a story for a music video for this song. {
            additional_user_prompt if additional_user_prompt else ''} The story should be engaging
            and visually appealing, with a focus on the emotions conveyed by the music and lyrics
            ."},
    ]},
]
```

Block 5. LALM-generated video script

In the music video the protagonist is a young woman named Elina who dreams of becoming a professional dancer. She faces numerous obstacles and rejections but remains determined to pursue her passion. One day, while practicing her routine alone in an empty studio, she hears the track playing in the background. The music Empowers her, and she feels a surge of confidence. As the bass drops, she begins to dance wildly, letting go of all her worries. The music's message of perseverance and self-empowerment resonates deeply with her, and she knows that she has found her calling. The video captures her journey from despair to triumph, showcasing her growth and resilience through the power of dance and music.

7. Exploratory Generative Pipeline Evaluation: Survey Items

In the following, we present the survey items used for the preliminary evaluation described in [subsection 4.1](#). All items were rated on a 7-point Likert scale: where 1 = Strongly Disagree, and 7 = Strongly Agree.

Storytelling

1. The video featured consistent characters.
2. I felt connected to at least one of the characters.
3. The narrative had a clear beginning, middle, and end.
4. The characters had clear goals or challenges.
5. The progression of the story felt motivated by the music.
6. The characters' actions were easy to follow.
7. The storyline was easy to understand.

Open Text Box: *What was your impression of the story and character development? What stood out to you—whether positive or negative?*

Visual Impression

1. The visual style of the video was aesthetically appealing.
2. The visual style of the video was distracting or confusing. (*reverse-coded*)
3. The visual elements supported the mood of the music.
4. The scene changes in the video matched the rhythm and structure of the music.
5. The visual setting and mood were appropriate for the music.

Open Text Box: *What was your impression of the visual style of the video? What stood out to you—whether positive or negative?*

Video Cutting / Transitions

1. Transitions between scenes felt natural.
2. I noticed cuts that felt abrupt or distracting. (*reverse-coded*)
3. The timing of scene changes matched the rhythm of the music.
4. The timing of scene changes matched the narrative flow.
5. The pacing of the video supported the unfolding of the story.
6. Some scene changes felt disconnected. (*reverse-coded*)

Open Text Box: *What was your impression of the pacing and transitions between scenes? What stood out to you—whether positive or negative?*

Emotional Consistency

1. The emotions conveyed in the video matched the emotional tone of the music.
2. The emotional tone of the video changed abruptly or felt disjointed. (*reverse-coded*)
3. I felt emotionally engaged with the video throughout.
4. The emotional tone remained consistent across different scenes.

Open Text Box: *What was your impression of the emotional consistency of the video? What stood out to you—whether positive or negative?*

Overall Video Impression

1. I enjoyed the video.
2. The video matched the emotional tone of the music.
3. The video felt repetitive or boring at times. (*reverse-coded*)
4. I found the video creatively engaging.
5. I would share or recommend this video.
6. I took the video seriously as a creative work.

Open Text Box: *Please reflect on the video's overall quality. What aspects worked well? Where did it fall short or feel inconsistent with the music or prior descriptions?*