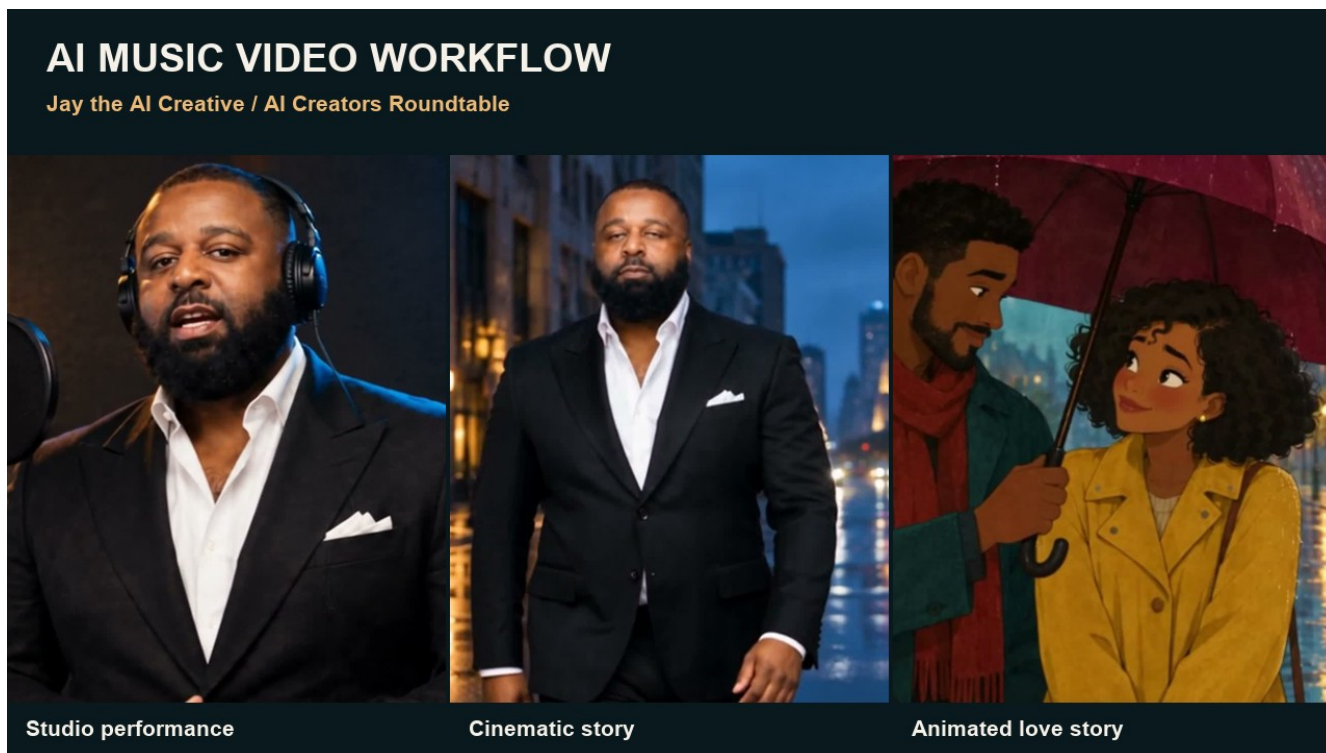


How to make an AI music video with Suno and Higgsfield

Jay the AI Creative | October 3, 2026



You can make an AI music video with a song and a clear photo. The useful part is knowing what to generate, how to keep your original song intact, and how to judge the export.

I tested three approaches: a studio rap promo, a cinematic story with the same character in different locations, and a short animated love story. This guide takes you through the inputs, prompts, edit, and quality checks. We also learned that matching a generated preview does not prove you have preserved the original song.

The tools handle production work. You still choose the song, direct the scenes, and decide what is good enough to publish.

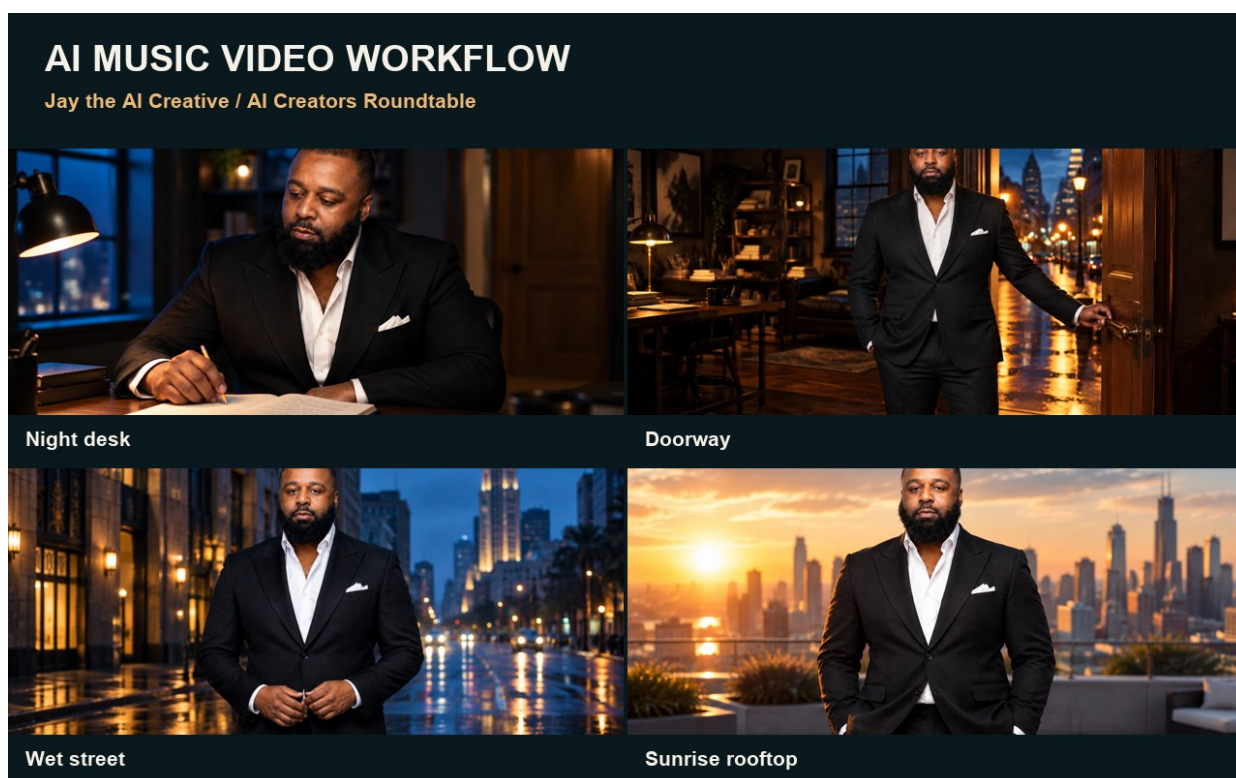
Presentation disclosure: the companion tutorial is presented by my digital avatar, Jay the AI Creative, using HeyGen Avatar III and my selected voice. Its job is to help creators and solo business owners use AI to buy back time and increase productivity. I direct the work; the avatar presentation is AI generated. The demo visuals are generated, and the animated couple is fictional.

The three examples

These are three separate examples with different recordings. You can reuse the workflow with your own song.

Example	What we made	How the audio works
Set the Frame	A roughly 45-second studio rap promo using my suit character	The reviewed standalone edit keeps the complete original song under the repaired performance
From the Spark	A full 48.4-second cinematic video: desk, doorway, street, rooftop	The original song plays continuously under silent story scenes
Still Here	A 25-second animated love story across three stages of a couple's life	An original ElevenLabs piano score plays under the scenes

[Watch the studio promo with the original song.](#) [Watch the full cinematic demo.](#) [Watch Still Here.](#)



The same suit character in the four From the Spark locations.

What you need before you start

- A song you can use, with the downloaded audio and lyrics saved locally.
- A clear photo of yourself, or a character image you have permission to use.
- Suno if you need to create the song, and Higgsfield for the visuals.
- An editor that lets you trim picture around a continuous song and check mouth timing against that recording.
- A folder for prompts, references, source clips, final exports, and review notes.

Start with a small test. A short song and one reference will teach you more than a long shot list you have not tested.

Choose the format first

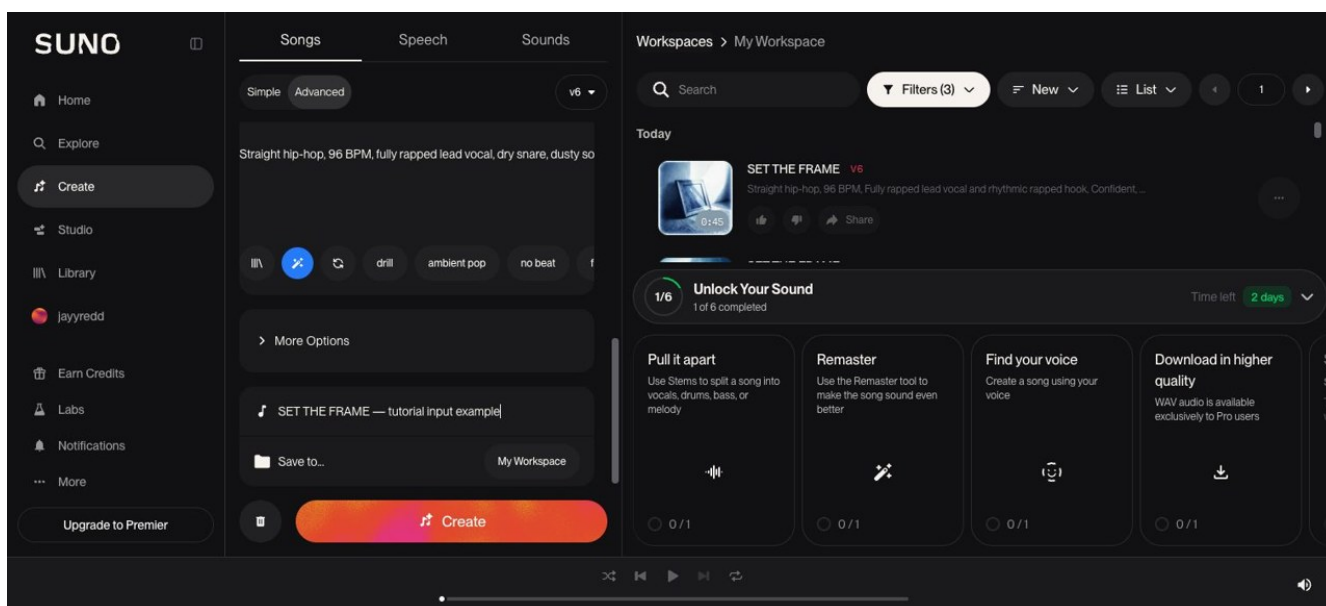
For an artist promo, try a performance to camera. For a brand film or artist visualizer, let a character move through a story while the song plays. For an anniversary or sentimental film, a short illustrated story can work with an instrumental score.

The choice affects the whole edit. A performance needs mouth timing as well as a song that stays intact. A narrative needs visual continuity around the complete recording.

Step 1: Create a short song in Suno

Write the lyrics separately from the style prompt. Our studio test used **hook** → **verse** → **hook**, with fully rapped vocals and a short arrangement.

In Suno's creation screen, use the advanced/custom lyrics controls available in your account. Put the lyrics in the lyrics field and the sound description in the styles field. Interface labels and models change, so use the controls currently shown to you.



Actual Suno creation screen with a prepared tutorial input example.

This screenshot is a prepared input example captured for the tutorial. It is not a screenshot of the original submission.

Copyable style prompt

Straight hip-hop, 96 BPM. Fully rapped lead vocal and rhythmic rapped hook. Confident, conversational delivery with sharp diction and a relaxed pocket. Hard kick, dry cracking snare, deep bass, dusty chopped soul textures, sparse piano stabs, subtle vinyl grit. Space for the bars, restrained ad-libs. Clean lyrics. Short track: four-bar hook, eight-bar verse, four-bar hook. Vocals enter on the first beat. End with a hard drum stop.

Generate a few candidates and listen for three things: words you can understand, a hook you remember, and an ending that feels finished. Choose the actual recording before planning the shots. A prompt asking for 45 seconds does not guarantee that exact runtime.

Save the audio, lyrics, prompt, selected version, and duration together. The complete original Set the Frame recording is **44.840 seconds**. From the Spark is 48.4 seconds. Use the original recording to set the timeline and keep its ending. A player rounding 44.840 seconds to 45 does not mean another player showing 44 has cut a second.

For monetized work, check your specific song's license under the current rules. Suno's help center says a later subscription does not grant retroactive commercial rights by default, while its newer downloads FAQ describes updated rights tied to paid downloads. Confirm the license attached to your track rather than assuming every song in your library has the same rights. See [Suno's retroactive-rights explanation](#), its [downloads and terms update](#), and [current terms](#).

Step 2: Build a character reference you can judge

Use a photo with visible features and clean lighting. I selected the black suit portrait and reused that identity for the booth and cinematic locations.

Generate a still for each location before asking for animation. Describe what stays the same first, then describe the setting and framing.

Booth still prompt

Preserve the reference person's facial features, skin tone, hair, beard, and build. Keep the black suit, open white shirt, and white pocket square. Place him in a recording booth wearing studio headphones. Warm amber practical light, blue rim light, realistic cinematic photography. Medium framing. Put the microphone off to one side so his mouth remains visible. Natural hands. One person. No text, logos, or watermarks.



The approved booth reference with the face and mouth visible.

Check identity, clothes, hands, and framing. If the still looks wrong, fix it now. Animation will carry those problems into more frames.

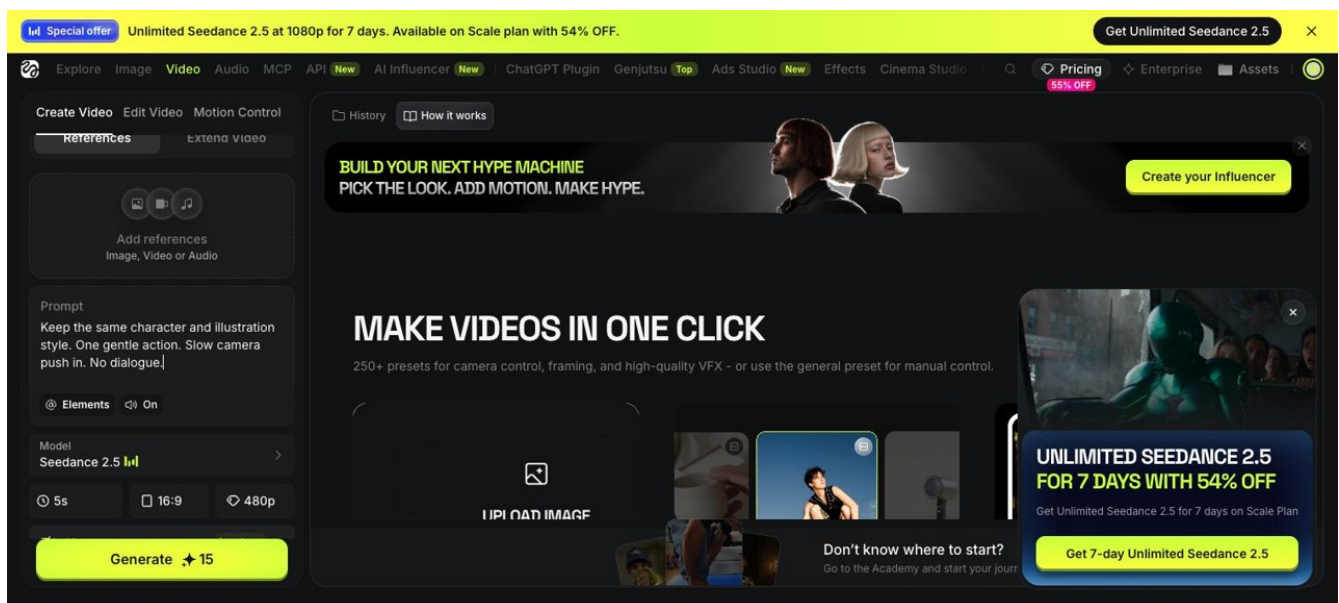
Step 3A: Make a studio performance in Higgsfield

Treat the booth footage and the finished song as separate inputs. The original recording is the master. Check the mouth against that exact recording before building the rest of the video.

Generate the booth footage

1. Open Higgsfield's video generation screen.
2. Choose a model and mode that support your reference image and the inputs you need. Our earlier booth takes used **Seedance 2.5** with image and audio references.
3. Upload the approved booth still. If the model accepts guide audio, use an excerpt from the original song with its exact start and end saved.
4. Set a supported duration, read the current credit estimate, and request restrained movement with a clear face.
5. Generate one short test. Listen to the returned sound and compare the mouth with the original recording. A model accepting audio as a reference does not prove it will return the same song.
6. Save the source and prompt. Continue only when you have a usable performance or a lip-sync method tested against your song.

Higgsfield's [Seedance help page](#) explains the generation controls, and its [official media-input reference documents](#) model-specific inputs. Check the selected mode in your account.



Actual Higgsfield video controls with Seedance 2.5 and the 480p setting.

This capture shows the earlier Seedance generation controls. The dedicated lip-sync step below uses a different model. We used the connected Higgsfield tools for these jobs.

Performance direction

Use the same man and booth from the reference. Perform the supplied rap audio with clear facial visibility and restrained natural head movement. Keep the suit, headphones, lighting, and camera framing consistent. Keep the microphone away from the mouth. No scene changes or extra people.

Our original excerpts cover **0-15, 15-30, and 30-44.840 seconds**. They preserve the full song. Keep those exact boundaries in your records; do not guess them from the lyrics or the rounded player label.

We requested **480p** for the booth footage and enlarged locally to 1080p. The lower resolution did not reduce the quoted cost in our generation test. Treat that as a recorded result from this account and model, not a pricing rule.

Drive the lips with the original song

If the generated mouth follows a changed version of the song, use a dedicated lip-sync pass with the finished recording. Higgsfield lists **Sync Lipsync 3** as a video-to-video lip-sync model. Sync's documentation supports song lip sync and recommends sync-3 for that use. The result still needs your review. See [Higgsfield's lip-sync guide](#) and [Sync's song guidance](#).

1. Prepare the clean performance video with the face visible. Add equipment cutaways, titles, and overlays after the lip-sync pass.
2. Supply that video and the **complete original song**. Keep the original recording saved separately for the final edit.
3. Check what the tool does when the two inputs have different lengths. In this booth test, the picture is 44.916667 seconds and the song is 44.840 seconds. We selected `silence` mode: the documented behavior keeps the video and adds silence after the shorter audio ends. See [Sync's duration modes](#).
4. Inspect the returned video's dimensions, frame rate, duration, face, and mouth timing at the beginning, middle, ending, and shot changes. Review it at normal speed with the original song playing.
5. Build the standalone edit around the complete original recording. Compare its sound with the original and check the complete ending before using it in the tutorial.

A clean vocal stem can help when the instrumental interferes with lip sync. If you use one, verify that its timing matches the full song, and keep the full original mix as the final soundtrack. Sync recommends isolated vocals in its [song guidance](#). This demo uses the complete original mix.

Our revised studio film: we used Sync Lipsync 3 on the existing three booth takes with the complete original MP3, then built the standalone edit with that original recording. Its full **44.840 seconds** is preserved from time zero. All 26 distributed sound comparisons measured zero added offset. The picture retains its native **24 fps and 1,078 frames**.

We then measured visible mouth movement against the original soundtrack using the [authors' SyncNet implementation](#). Known 200-millisecond sound advances and delays confirmed which way its offset estimates point. Before correction, it estimated the opening mouth movement was **80 milliseconds late**, the middle was aligned, and the ending was **120 milliseconds late**. We advanced only the first take by **two native frames (83.3 milliseconds)** and the last by **three (125 milliseconds)**. The middle and the complete soundtrack stayed unchanged. These are measured adjustments for this film, not settings to copy into every performance.

After correction, all three performance sections favored **zero offset** in the diagnostic. All **20 three-second windows with sufficient diagnostic confidence** favored an offset within **40 milliseconds** either way; **17 favored zero**. Three early windows were weak or ambiguous and did not justify another correction. An independent review recomputed the results and passed this measured mouth/audio check.

That pass has limits. The diagnostic compares timing in 40-millisecond steps, while each native picture frame lasts about 41.7 milliseconds. A zero estimate does not certify every phoneme. Rap over a full musical mix, facial hair, and source resolution can affect the measurement. This review did not include uninterrupted human audiovisual playback. Review the complete result with sound and decide whether the performance works for you.

Smooth the joins, then review the standalone film

Independent takes can jump in pose or gaze. The reviewed standalone edit uses moving microphone and headphone cutaways at **13-17 seconds** and **28-32 seconds**, with half-second dissolves at each edge. Those shots cover the take changes while the complete original song continues. The fades briefly layer equipment over the performer as they return to the booth view.

Edit and check the complete studio film before placing it in the tutorial. Technical checks establish that the file plays and retains the song. Editorial review checks the performance, identity, transitions, and framing in order across the film. Mouth accuracy needs its own listening and viewing check.

What our earlier lip-sync fix missed

The Higgsfield previews looked better synchronized than the first local export. In that first edit, we advanced the picture by **40, 50, and 125 milliseconds** while using the original MP3. The mouths led the words.

Removing those unverified picture advances and using each preview's embedded sound preserved the preview timing. It also replaced the song with generated sound that had different timing. The earlier export checks proved that the edit preserved that preview; they did not prove it preserved the original recording.

The release requirement is now clear: **keep the complete original song as the master and test the performance against it**. Repair or replace a performance that does not fit. Review the standalone film first, then check the actual tutorial export after integration.

Step 3B: Create a cinematic story over the original song

For From the Spark, the character does not sing or rap. He moves through four locations while the original song stays continuous.

Scene	Timeline	Direction
Night desk	0-12 seconds	Write, then put down the pencil
Doorway	12-24 seconds	Walk through the doorway
Wet street	24-36 seconds	Walk through the city; ease into a 12% closer view from 30-32 seconds
Sunrise rooftop	36-48.4 seconds	Look toward the sunrise

Use the same identity reference for every still. Then describe **one clear action** for each animation. Put the complete song on the edit timeline and trim the scenes around it.

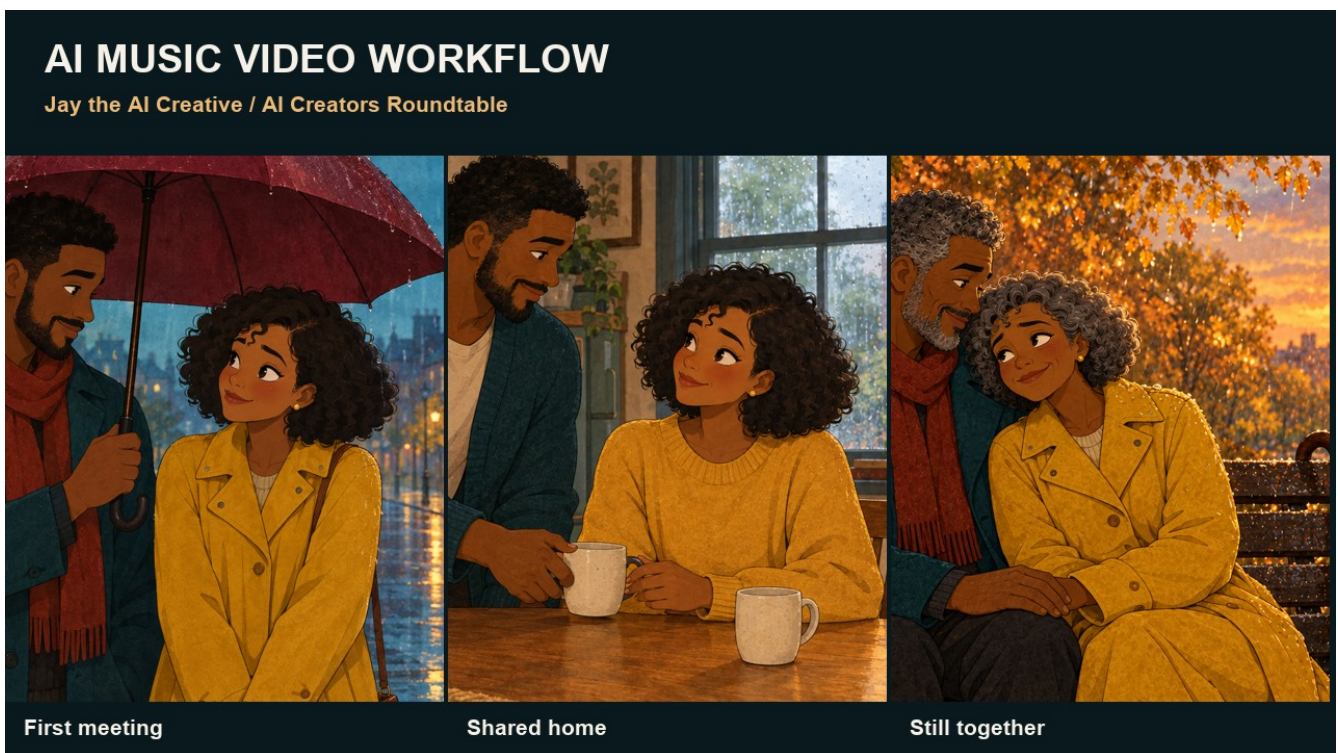
Reusable scene prompt

Preserve the same character, face, beard, build, and black suit.
Animate the approved location still as one continuous cinematic shot.
Action: [one specific action].
Camera: [one restrained movement or a locked shot].
Keep the character recognizable and centrally framed.
No speech, no singing, no lip-sync performance, no extra people.

This format is useful when the mood and story matter more than performing every lyric. It also avoids replacing the original song to match a generated mouth.

Step 3C: Make a short animated love story

Still Here follows a fictional Black couple through a rainy first meeting, a quiet kitchen, and growing old together. A burgundy umbrella connects the scenes.



The animated couple at their first meeting, at home, and later in life.

Start with a reference showing both people. Preserve their faces, skin tones, clothing colors, and illustration style across the next stills. For the older scene, specify gray hair and natural age lines while retaining their recognizable features.

Character and style direction

Hand-painted 2D illustration with warm watercolor textures and clean outlines.
A fictional Black couple: he has deep brown skin, short curly hair, a short beard, a teal jacket and rust scarf. She has medium brown skin, a rounded curly bob, a mustard coat and small gold earrings.
Keep both people recognizable throughout. Burgundy umbrella as a recurring object. Gentle expressions, natural hands, cinematic composition. No text.

Three scene actions

- **First meeting:** he gently tilts the umbrella to cover her; she smiles.
- **Shared home:** he slides a cream coffee mug toward her in a cozy kitchen.
- **Later life:** she rests her head on his shoulder on a sunset park bench.

Each generated source was ten seconds. The final edit uses ten seconds, ten seconds, and the first five seconds of the last shot. The last generation pushed the camera too low later in the shot, so we cut before it lost the faces. That gave us a stronger **25-second** film.

The original score was generated with ElevenLabs: intimate felt piano, a simple warm motif, light cello, no drums, no vocals. Its ending fades to fit the shorter edit. A small gesture and a recurring object can carry the story without complicated dialogue.

Step 4: Assemble the edit and teach with the visuals

Put the complete original recording on the timeline first. For a performance, use footage whose mouth timing has been checked against that recording. Keep tested video offsets and speed consistent as you trim the picture around the song.

Use crop changes to create another composition without generating another shot. Keep useful context on screen long enough to read it. For this tutorial, the talking-head crop holds are at least five seconds, while the voice remains continuous within each chapter.

The companion presentation uses **HeyGen Avatar III**, real browser screenshots, and **Hyperframes** for titles, prompts, timeline diagrams, captions, and layout changes. The screens distinguish actual interfaces from prepared examples and diagrams reconstructed from our edit records.

Make the widescreen cut first, then inspect the vertical version separately. A centered subject helps, but you still need to check where the face moves. Enlarging 480p to 1080p changes dimensions; it does not restore detail that the generation never made.

Step 5: Run three quality checks

Round	Check	What it catches
File and song	Decode the complete export; compare its sound with the original recording, including the start and full ending; verify dimensions and frame rate	Broken files, changed or missing song material, black gaps, accidental freezes
Timing	Compare the original song with the export at the opening, joins, and ending; watch the mouth against it	Added offsets, accumulating drift, generated lips following different timing
Picture	Inspect identity, hands, crop, readable text, continuity, and transitions across the complete film	Visual problems that file metadata and sound comparisons cannot identify

The earlier studio test compared 42 sound windows with the generated preview and found zero added offset. That result describes preview preservation. It cannot be reused as proof that the new edit

preserves the original song or that its new lips are accurate. The cinematic test compared 46 windows with its uninterrupted original song at zero offset.

For the studio revision, the full 44.840-second original recording is preserved and all 26 distributed sound windows measured zero added offset. The separate mouth/audio diagnostic favored zero for all three performance sections, with the limits described above. Check the actual tutorial export again after integration. Record the compared source, result, and scope for each round, and review the complete result with sound before publishing. Sound comparisons and sampled frames do not certify every mouth movement.

If something looks wrong

Problem	First useful check
Mouths lead the words	Confirm the original song is intact, check unverified picture offsets, then review the mouth against that exact recording; repair or replace the performance if needed
Face changes between locations	Revisit the approved identity reference and regenerate the still before the animation
Hands or props behave strangely	Simplify to one action; inspect the source and trim unusable portions
Vertical crop cuts off the face	Review the full movement in the vertical framing
1080p still looks soft	Check source resolution; ordinary enlargement cannot recover missing detail

Your first test

Make one clear reference and one short clip. Save the prompt, source, and original recording. Check the mouth and export against that song, including its ending. When that works, build the rest.

For another workflow, see [put yourself in a music video with AI](#), the [AI video producer walkthrough](#), and [six creator tasks for your AI assistant](#).

AI handles the busywork. You keep the taste.