

05 · Video

AI Video Planning Template

A template for planning a short video before any generation: audience, hook, message, duration, shot list, sound, transitions, reference assets and export needs, with a completed 30-second storyboard.

For: Anyone producing short videos for a brand or association with AI video and voice tools.

Level: Intermediate

Version 1.0 · 2026-09-21

Why plan before you generate

What burns the most time and credit in AI video is generating without a plan: a beautiful shot that does not fit what follows, a voiceover longer than the pictures, an idea that changes halfway through. Planning turns the video from a random experiment into a chain of small, clear decisions, and makes every generation serve a specific place in the video.

Rule one: script first. A 30-second video is roughly 4 to 5 shots, each with what we see plus what we hear. At the time of writing (September 2026), video generators mostly produce short clips, and results vary a great deal from one tool to another. So we plan short shots and assemble them in an editor rather than asking for a whole video from one prompt.

A second rule matters just as much: **a good short shot beats a bad long video.** If you cannot produce five good shots, three good shots and a 15-second video beat 30 seconds with two distorted shots.

Sections of the planning template

1. Foundation: goal, audience, platform

Start with three questions: what do we want the viewer to do after the video? Who exactly are they? Where will they watch it? The platform sets the aspect ratio, a sensible length and the viewing context: vertical video on a phone is usually watched while scrolling fast, and many people watch without sound, so the message must be understood from the picture and on-screen text alone.

2. The single message

A short video carries one message only. Write it in a sentence of no more than 12 words. If you need "and" twice to write it, you have two videos. The message is what the viewer should remember if they forget everything else.

3. The hook: the first three seconds

The hook is what stops the thumb from scrolling. It comes first in the picture, then in the words. Patterns that work well:

- **An unexpected image or strong movement:** steam rising from a cup in an extreme close-up, or a hand slowly turning something over.
- **A problem the viewer lives with:** "Mornings always start in a rush?" with an image that shows it.
- **The result before the method:** show the satisfying end first, then go back to explain.
- **A specific fact:** a real, checkable sentence about your product or service, not a general claim.

Avoid opening with the logo or a long greeting; these are the first three and most expensive seconds.

4. Duration and word budget

Set the total length, then divide it among the shots. For voiceover, a working rule: a 30-second video carries no more than about 60 spoken words, with pauses that let the picture breathe. Read the script aloud with a timer before producing the voice; it exposes a long script immediately.

5. The shot list

Each shot gets a row with: number, timing, what we see (a precise visual description that can serve as an image prompt), camera movement, what we hear (voice, music, effects), on-screen text, the source of the shot (generated image then animated, text-to-video, real footage, screen capture) and the transition to the next shot.

Image to video: animating a still image (generated or real) gives more control over shape and colour than generating video directly from text. So the preferred route is: generate one image per shot in a unified style (with the same fixed style sentence), then animate it with a simple camera-movement description.

6. Sound

- **Voiceover:** Modern Standard Arabic or dialect? Try both if you can and choose what is closest to your audience. Control pace and pauses with punctuation in the script.
- **Music:** describe the mood and tempo (calm, morning, mid-tempo) and confirm you have the right to use it commercially.
- **Sound effects:** the sound of coffee pouring or a door opening adds realism at little cost.
- **Levels:** speech is always clearer than music; keep the music clearly under the voice and lift it where there is no speech.
- **On-screen captions:** essential, because a large share of viewing happens without sound. Check automatic transcription letter by letter, especially in Arabic.

7. Transitions

In short video, the straight cut is the main transition and usually the best. Use others deliberately: a **match cut** (a shape or movement continues from one shot into the next), a **cut on motion** (cutting during a camera move so the change feels natural), and a **fade** for a calm ending. Avoid lots of showy transitions;

they draw attention to the edit rather than the message. Plan the transition when you write the shots: if shot one ends on a push-in to the cup, let shot two start on something round or with movement in the same direction.

8. Reference assets

Before generating, gather everything the video needs: real product photos, a high-quality logo, brand colours and fonts, the style sentence, a character reference if there is one, licensed music, and any real footage. If the video includes a real person's face or voice, their explicit written consent is a precondition for any use, and nobody's voice is cloned without written permission.

9. Export requirements

- **Aspect ratio:** 9:16 for stories and vertical short video, 1:1 or 4:5 for feed posts, 16:9 for websites and YouTube. If you need several ratios, plan compositions so the subject stays in the safe centre.
- **Safe zones:** in vertical video the app interface covers the top, bottom and side of the screen; keep text and logo away from the edges.
- **Resolution and frame rate:** make them consistent across all shots before assembly, and export at the resolution the platform asks for.
- **Captions:** burned into the picture or a separate file? Some platforms prefer the separate file.
- **Cover:** a clear frame or cover image that represents the video.
- **File naming:** a name that includes project, ratio and version, such as `daraj-breakfast-9x16-v2`.

The blank template

Video planning template

Project: [] Date: [] Owner: [role]
Goal (what the viewer does next): []
Audience: [who exactly]
Platform and ratio: [] Total length: [] seconds
Single message (max 12 words): []
Hook (first 3 seconds): picture [] / words or on-screen text []
Call to action: []

Sound:

- Voiceover: [MSA / dialect], [tone], word limit: []
- Music: [mood and tempo], licence: []
- Sound effects: []
- Captions: [burned in / separate file], checked by: []

Reference assets: [product photos, logo, colours, font, style sentence, character reference]

Consents: [real faces or voices? written consent: on file / not needed]

Fixed style sentence: []

Export: ratios [] / resolution [] / cover [] / file name []

Success criterion: [how we know the video did its job]

Shot list

#	Time	What we see	Camera	What we hear	On-screen text	Source	Transition
1	0-3	[]	[]	[]	[]	[]	[]
2	3-[]	[]	[]	[]	[]	[]	[]
3	[]	[]	[]	[]	[]	[]	[]
4	[]	[]	[]	[]	[]	[]	[]
5	[]	[]	[]	[]	[]	[]	[]

The video's single message

The hook: what do we see and hear in the first three seconds?

The full voiceover script (read it aloud with a timer)

Assets to gather before starting, and consents required

A helper prompt for the script

You can ask a chat model for a first draft of the script in table form, then improve it yourself:

Script prompt

You are a social media ad writer.
Write a [length]-second script for [product or service] aimed at [audience]. Single message: [message]. Call to action: [CTA].
Format: a table of [number] shots with columns: Shot no. | What we see (a precise visual description usable as an image prompt) | Camera movement | Voiceover (one sentence) | Duration in seconds.
Constraints: open with a visual hook in the first 3 seconds and close with the call to action. No more than [number] spoken words in total. Make no claims beyond this real information: [real product information].

The workflow from plan to final file

1. **Write the script and shot list** and read the voiceover aloud with a timer.
2. **Generate one image per shot** with the same style sentence, and check product fidelity and consistency across images before animating.
3. **Animate each image** with one simple camera-movement description, generating two or three versions per shot to choose from.
4. **Produce the voiceover**, try two voices, and choose with someone from the target audience if possible.
5. **Add music and effects** and set the levels.
6. **Assemble in a video editor**: cut the shots to the voiceover, not the other way round, then add text and captions.
7. **Review**: watch once without sound, once on a phone, and once through the eyes of someone who has not seen the plan.

8. **Export** at the required ratios and resolution, name files clearly, and save the plan with the prompts used.

A completed storyboard: a 30-second video

Illustrative example

Project: Daraj Café near the university, launching a breakfast menu. **Goal:** get students and office workers to drop in and try breakfast before 10 am. **Audience:** university students and employees who start their day early. **Platform and ratio:** stories and short vertical video, 9:16. **Length:** 30 seconds.

Single message: a quick, warm breakfast before your lecture or work, ready in minutes. **Call to action:** "Drop by before ten." **Style sentence:** "Realistic food photography in warm morning window light, wood and cream tones with touches of orange, shallow depth of field, a calm and welcoming feel, no text or logos."

#	Time	What we see	Camera	What we hear	On-screen text	Source and transition
1	0-3	Extreme close-up: coffee poured into a white porcelain cup, steam rising in the window light.	Static with a very slow push-in	Sound of pouring, then voiceover: "Mornings start in a rush?"	"Mornings start in a rush?"	Generated image, then animated; straight cut
2	3-9	A wooden table by a window: a za'atar croissant, a plate of labneh with olive oil, a glass of orange juice.	Slow lateral move right to left	"We have a warm breakfast ready before your lecture or work."	"The new breakfast menu"	Generated from a photo of the real dishes; cut on motion
3	9-16	Two hands place a brown paper bag with the café's sticker on the counter, a takeaway coffee beside it.	Gentle push-in toward the bag	"Take it with you in a bag, or sit for a few minutes with your coffee."	"Grab and go, or stay"	Real phone footage (real sticker); straight cut
4	16-24	Wide shot of the café in the morning: two tables with people whose faces are indistinct, warm light, a menu board on the wall.	Slow dolly forward	"From seven thirty to ten, every day except Sunday."	"7:30 - 10:00 except Sunday"	Real footage of the café; match cut on motion
5	24-30	The cup from shot one on the table, the steam lighter, the logo in the space above.	Static	"Daraj Café. Drop by before ten." Then music alone.	Logo + "Drop by before ten"	Generated image with gentle animation; fade out

Full voiceover (about 40 words in Arabic): "Mornings start in a rush? We have a warm breakfast ready before your lecture or work. Take it with you in a bag, or sit for a few minutes with your coffee. From seven thirty to ten, every day except Sunday. Daraj Café. Drop by before ten."

Sound: voiceover in light Modern Standard Arabic with a friendly tone; calm, mid-tempo morning music licensed for commercial use, kept low under the voice and lifted for the last two seconds; a pouring sound effect in shot one.

Assets: real photos of the three dishes (the reference for food fidelity), the real bag sticker, the logo on a transparent background, the brand font. The people in shot four are customers who gave written consent to appear, with faces indistinct, or the shot is replaced with the empty space.

Export: 9:16 as the main version, plus a 4:5 feed version checked for safe zones; burned-in captions because most viewing is on phones; a cover taken from shot two; file name `daraj-breakfast-9x16-v1`.

Success criterion: a viewer without sound understands what is new and when to come; the food matches what will actually be served; no generated text inside images; times match the café's real opening hours.

Responsibility and ethics

- No deepfakes: no real person's face or voice without explicit, documented consent.
- The food or product in the video must match what the customer will receive.
- Disclose generated content where your audience expects it or where the platform's or your organisation's policy requires it.
- Confirm that music and fonts are licensed for every commercial use.

Common mistakes

- **Mistake:** asking for a full one-minute video from a single prompt. **Fix:** split it into short shots and assemble them in an editor.
- **Mistake:** generating video directly from text and getting inconsistent shots. **Fix:** start from one image per shot with a fixed style sentence, then animate.
- **Mistake:** a voiceover longer than the video. **Fix:** set a word budget and read the script aloud with a timer before production.
- **Mistake:** music drowning the speech. **Fix:** keep music clearly under the voice and listen on an ordinary phone speaker.
- **Mistake:** forgetting captions or publishing an automatic transcript with errors. **Fix:** add captions and check them letter by letter.
- **Mistake:** describing several camera moves in one shot. **Fix:** one simple move per shot; compound moves increase distortion.
- **Mistake:** placing text near the edges of vertical video. **Fix:** keep text and logo inside the safe zone.

Checklist before export

- ☐ The single message is clear to someone watching without sound.
- ☐ The hook lands in the first three seconds, and the logo is not the first thing shown.
- ☐ The voiceover is within the word budget and in sync with the shots.
- ☐ All shots share a consistent style and colours.

- ☐ The product or food matches the real thing in every shot.
- ☐ No distortions in hands, faces or objects, and no generated text inside images.
- ☐ Speech is clearer than music, and the music is licensed.
- ☐ Captions are checked and match the speech.
- ☐ Text and logo sit inside the safe zone in every exported ratio.
- ☐ Written consents are on file for any real face or voice.
- ☐ Files are clearly named, and the plan and prompts are saved.

About the author — Mohamad Charafeddine is an entrepreneur from Tripoli, Lebanon: founder of the Passion digital marketing agency, president of Takaful Youth Association, and founder of digital platforms. He trains individuals and teams in applied AI and marketing.

Examples in this file are fictional. Free to use in your work and organisation; credit the source if you republish.
mohamadcharafeddine.com/en/resources/ai-video-planning/