

08 · Video

From Idea to Short Video

Nine stages for making a short AI video: brief, script, storyboard, generation, selection, editing, captions, review and delivery, with a full example.

For: Anyone who wants to produce a short video (15 to 45 seconds) for a brand or organisation with AI tools, in an organised and repeatable way.

Level: Intermediate Version 1.0 · 2026-09-21

AI-generated short videos usually fail for one reason: the maker starts from the tool rather than the idea. They write a long description of a whole video, hit "generate", then spend hours chasing a result they cannot define. This guide reverses the order. The idea becomes a brief, the brief becomes a script split into shots, each shot is generated and reviewed on its own, and the video is assembled in an editor. That way every flaw stays small and fixable instead of sinking the whole piece.

One line to remember: a good short shot beats a bad long video. At the time of writing (September 2026), tools produce clips of a few seconds at a time, so we plan the video as shots from the start.

The whole picture: nine stages

#	Stage	Output	Who decides
1	Brief	One page: goal, audience, message, length, platform	Business owner or account lead
2	Script	Shot table: what we see, what we hear, duration	Copywriter
3	Storyboard	One reference frame per shot plus camera movement	Director or designer
4	Generation	Several takes per shot, named and saved	Producer
5	Selection	The best take for each shot, against written criteria	Producer with a reviewer
6	Editing	An assembled cut: shots, voice, music, transitions	Editor
7	Captions and on-screen text	Synced Arabic captions, title and call to action	Editor with copywriter
8	Review	Findings log and decision: publish or revise	A reviewer who was not in production
9	Delivery	Files in required sizes, post copy, source files	Producer

In a small team one person may hold every role except review; review always needs a second pair of eyes.

Stage 1: The brief

The brief answers one question: what should the viewer do, feel or know after the video? If you cannot answer in a sentence, the video is not ready for production.

- **Goal:** one action, as measurable as possible (order, visit, sign up, share).
- **Audience:** who they are, where they watch, what they already know.
- **The single message:** the sentence the viewer remembers if they forget everything else.
- **Length, platform and aspect ratio:** for example 30 seconds, vertical 9:16.
- **Available assets:** real product photos, logo, colours, font, approved voice.
- **Red lines:** what must not appear (real faces, health claims, unconfirmed prices).

Short video brief template

Brand / organisation: [..]
Goal (one action): [..]
Audience: [who, where they watch, what they know]
Single message: [one sentence]
Length: [15 / 30 / 45 seconds] Platform and ratio: [.. / 9:16]
Tone: [warm / energetic / calm / formal] Language or dialect: [..]
Available assets: [product photos, logo, colours, font, music]
Call to action: [exact wording]
Red lines: [..]
Due date: [..] Reviewer: [name]

Stage 2: The script

Write the script as a table, not a paragraph. A practical starting rule: 30 seconds is roughly four shots of about seven seconds, and each shot has two parallel columns: what we see and what we hear. Open with a hook in the first three seconds and close with a clear call to action. Watch the spoken word count: comfortable Arabic speech needs time, so a 30-second video carries no more than about sixty spoken words.

You can ask a chat model for a first draft, then edit it yourself:

Script draft prompt

You are a social media copywriter.
Write a [length]-second script for [product or service] aimed at [audience].
Goal: [desired action]. Single message: [sentence].
Format: a table of [4] shots with columns:
Shot # | What we see (precise visual, usable as an image prompt) |
What we hear (one Arabic sentence) | Duration
Open with a hook in the first 3 seconds and close with a clear call to action:
[CTA wording]. Total spoken words no more than [60]. Tone: [..].
Language: [MSA / dialect]. Do not mention prices or claims not in this brief.

Stage 3: The storyboard

The storyboard turns each script line into one still image: the first frame of the shot. These frames are your generation references, and they are the cheapest place to discover that a shot does not work. For each shot, write:

- **The full image prompt:** subject, pose and setting, style, lighting, angle and lens, colours, aspect ratio.
- **A fixed style sentence:** one sentence pasted at the end of every prompt (style, palette, "no text, no logos") so the shots look like one video.
- **Camera movement:** only one movement per shot (slow push-in, lateral slide, locked-off with movement inside the frame).
- **What must not change:** the product, the character's clothes, the light direction.

For physical products, start from a real photo whenever possible: animating a real or generated image gives far more control than generating video straight from text.

Stage 4: Generation

Generate shot by shot, not the whole video. For each shot: the reference frame from the storyboard, then a short, specific motion prompt.

Shot motion prompt

The attached image moves: [one camera move, e.g. slow push-in toward the product].
Movement inside the frame: [e.g. steam rising, a hand slowly placing the plate].
Must not change: [product, colours, lighting, clothing].
No text, no logos, no extra people.
Duration: [5–7] seconds. Aspect ratio: [9:16].

- Generate two to four takes per shot, and change only one element between takes so you learn its effect.
- Name every file as soon as you download it, with shot and take numbers (`shot2-take3.mp4`), and save the prompt in a text file next to it.

- Voiceover: write the script with deliberate punctuation; commas and full stops control the pauses. Try at least two voices and choose by ear, not by description.
- Music: describe mood, rhythm and length, and confirm the tool's terms allow commercial use.
- Real voices or faces: no cloning and no avatar of a real person without written consent.

Stage 5: Selection

Choosing by eye alone favours the "prettiest" take rather than the right one. Write your criteria before watching the takes, then score each one:

Criterion	Question
Fidelity	Does the product match the real one in shape and colour?
Safety	Is it free of extra fingers, distorted faces, invented text or logos?
Continuity	Does it match the shots before and after (light, colours, clothes)?
Motion	Is the movement smooth, and does it end on a usable cut point?
Purpose	Does it say what this shot must say in the script?

If no take works after four attempts, the problem is usually the reference frame or an overly complex movement. Simplify the shot rather than keep generating.

Stage 6: Editing

Tools produce shots; the video is assembled in an editor. A suggested order:

1. Lay the voiceover on the timeline first; it is the backbone of the timing.
2. Place the shots over it, trimming each to start and end on a clean movement.
3. Add music under the voice and duck it during speech (a common starting point: voice at full level, music at roughly a fifth, then adjust by ear).
4. Use simple transitions; a straight cut is enough most of the time.
5. Add the logo and the call to action at the end using brand fonts.
6. Export a draft and watch it on a phone before any further change.

Stage 7: Captions and on-screen text

Many viewers watch without sound, so captions are part of the message, not decoration.

- Use the editor's auto-captions as a starting point only, then correct every word by hand; tools get names and dialect wrong.
- One or two lines at a time, few words, on screen long enough to read.
- Place text in the lower-middle third, away from app buttons and frame edges.
- Arabic text inside the image is added in the editor; do not ask the generator for it.

Stage 8: Review

A reviewer who was not in production watches on a phone, once with sound and once without, using a fixed checklist (see the resource "Pre-Publish Checklist for AI Images and Video"). The key checks: product fidelity, hands and faces, continuity between shots, correct Arabic text and numbers, pronunciation of names, music level, permissions. Each finding is logged with its timestamp and classed as blocker, fix, or next time.

Stage 9: Delivery

Delivery file list

- [] Final video in the main ratio (e.g. 9:16) — file name with version number
 - [] Extra ratios if requested (1:1, 16:9)
 - [] Cover image (first frame or a designed image)
 - [] Separate caption file if the platform supports it
 - [] Post copy + hashtags + call to action
 - [] Source folder: editor project, selected shots, voice, music
 - [] Prompt file: every image, motion and voice prompt
 - [] Permissions and licences (written consents, music licence)
 - [] Review log signed with the reviewer's name
-

How long does it take?

Time varies a great deal with team experience, tool and shot complexity, so we give no fixed figures. What can be said with confidence: stages 1 to 3 feel "slow" but shorten stages 4 and 5 considerably, because most of the time lost in generated video comes from random generation without a storyboard. Log your time on your first three videos to see where it really goes.

A complete example, start to finish

Illustrative example

Al-Zaytouna Bakery is a fictional bakery in Tripoli launching morning manakish delivery to nearby neighbourhoods. It wants a 30-second vertical video for Instagram and TikTok.

1. Brief

Goal	The viewer sends a WhatsApp message to order a first morning delivery
Audience	Families and office workers in nearby neighbourhoods, watching on their phones in the morning
Single message	Your manousheh arrives hot before you leave home
Length and ratio	30 seconds, 9:16

Tone and language	Warm and simple; voiceover in light Tripoli dialect, on-screen text in simple MSA
Available assets	Real phone photos of the manousheh and of the branded paper box, brand colours (olive green and cream), logo
Call to action	"Order on WhatsApp" with the phone number (typed in the editor and checked character by character)
Red lines	No real faces, no prices in the video, no promise of delivery within a set number of minutes

2. Script (after editing the model's draft)

#	What we see	What we hear (dialect, translated)	Duration
1	A phone alarm ringing on a kitchen table in blue dawn light, an empty tea glass	"Mornings start early... and no time for breakfast?"	6 s
2	A baker's hands stretching dough and spreading za'atar, a lit stone oven behind	"We bake from dawn."	7 s
3	A manousheh leaving the oven, steam rising, then placed in the green paper box	"Your manousheh, hot, folded in its box..."	8 s
4	The box on an apartment doorstep, a hand opens the door and picks it up (no face), then logo and CTA	"...and it reaches you before you leave home. Order on WhatsApp."	9 s

Total: 30 seconds and about 25 spoken words, which leaves the images room to breathe.

3. Storyboard

Fixed style sentence: *"Warm realistic photography, olive green, cream and bread-brown palette, soft morning light, shallow depth of field, no text, no logos, no faces, 9:16."*

- **Shot 1:** full image prompt for the table and alarm + style sentence. Movement: locked-off, only the phone vibrates.
- **Shot 2:** generated image of two hands stretching dough (no face) + style sentence. Movement: slow lateral slide.
- **Shot 3:** starts from a **real photo** of the manousheh and green box, because the product must match reality. Movement: slow push-in with steam.
- **Shot 4:** starts from a real photo of the box on a doorstep. Movement: locked-off, a hand enters from the right.

Must not change: the green box colour, the folded manousheh shape, light coming from the left.

4. Generation

- Shot 1: three takes. The first showed a phone screen with clear English text, so "phone screen off" was added to the prompt.
- Shot 2: two takes; the second is usable.
- Shot 3: four takes. In two, the manousheh turned into a pizza with melted cheese, so the prompt was simplified to "steam only, no change to the food".

- Shot 4: three takes; one showed a six-fingered hand.
- Voice: two dialect voices were tried and the warmer one chosen. A comma was added after "early" to lengthen the pause.
- Music: light oud and a calm rhythm, 30 seconds, after checking the tool's commercial-use terms.

5. Selection

Shot	Chosen	Why
1	Take 3	Screen off, convincing dawn light
2	Take 2	Natural hand movement, oven in the right colours
3	Take 4	Manousheh matches the real photo, natural steam
4	Take 1	Hand is correct, box in its real colour

6. Editing

The voiceover went down first, then the shots were trimmed to it. Shot 3 was cut one second early because the steam "melted" unnaturally at the end. Music was ducked under speech and lifted slightly for the last two seconds. Logo and number were added in the editor for the final three seconds.

7. Captions

Captions were auto-generated, then corrected by hand and written in simple MSA: "The morning starts early... no time for breakfast?", "We bake from dawn", "Your manousheh, hot in its box", "It reaches you before you leave home". They sit in the lower-middle third.

8. Review

A colleague outside production reviewed on her phone. Findings: the phone number at the end was missing a digit (blocker, corrected against the source); the shot 2 caption ran half a second ahead of the voice (fix); the word "dawn" was nearly lost under the music (fix). After revision, only those three spots were re-checked, and the decision was: publish.

9. Delivery

Delivered files: `zaytouna-delivery-9x16-v4.mp4`, a square version for the page, a cover image from shot 3, post copy with the call to action, the source folder, the prompt file and the review log.

Common mistakes

- **Mistake:** asking for a whole video in one prompt. **Fix:** split it into shots from the start, one prompt per shot.
- **Mistake:** generating video straight from text, with inconsistent results. **Fix:** start from an image, ideally a real product photo.
- **Mistake:** a script crammed with words. **Fix:** count spoken words against seconds and cut before you speed up the voice.

- **Mistake:** music louder than the voiceover. **Fix:** duck it under speech and listen on a phone's small speaker.
- **Mistake:** trusting auto-captions without correction. **Fix:** correct every line, especially names, numbers and dialect.
- **Mistake:** randomly named files and lost prompts. **Fix:** name each take on download and save its prompt next to it.
- **Mistake:** generating for hours on a single shot. **Fix:** after four failed takes, simplify the shot or change the reference image.

Completion checklist

- ☐ The brief fits on one page, and the single message fits in one sentence.
- ☐ The script is a shot table with what we see, what we hear and duration; spoken words are counted.
- ☐ Each shot has a reference frame and one camera move, and the style sentence is in every prompt.
- ☐ Every take is named and saved with its prompt.
- ☐ Selection used written criteria, not impressions.
- ☐ Editing started from the voice, and the music sits below speech.
- ☐ Captions are hand-corrected and placed in a safe area.
- ☐ A reviewer outside production checked on a phone and logged the decision.
- ☐ The delivery set is complete, including source, prompts and permissions.

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/idea-to-short-video/