AI Text-to-Video Prompting Cheat Sheet

Master the art of crafting precise, cinematic prompts for AI video generation. This comprehensive guide teaches you the language and techniques that transform vague ideas into polished, professional video outputs—no API knowledge required, just prompt craft.

Shot Blueprint and Structure

The foundation of exceptional AI video generation begins with architectural thinking. Before describing visual details, establish a clear structural framework that prevents the model from making arbitrary decisions. Think of your prompt as a blueprint where every element serves a purpose and builds toward a unified vision.

Start every prompt with a one-line logline that captures the essence of what the viewer should experience. This single sentence acts as your north star, preventing the common pitfall of meandering descriptions that confuse the model. For example: "A moment of quiet realization as a character confronts an empty room."

The three-part skeleton provides your organizational backbone: **Subject → Action → Setting**. Once this core is locked, layer in refinements: **Style → Camera → Lighting → Mood**. This hierarchy ensures the model prioritizes what matters most—the fundamental scene—before applying aesthetic flourishes.

Logline First

Single sentence capturing viewer's intended experience prevents rambling prompts

3-Part Skeleton

Subject → Action → Setting forms the core structure before adding details

"What's On Screen?"

Explicitly list visible elements to reduce model guesswork and ambiguity

Layer Ordering

Critical constraints early (subject, action), aesthetics later (style, color)

Two critical passes solidify your structure. The "What is on screen?" pass forces you to explicitly enumerate visible elements—people, props, weather conditions, signage—leaving nothing to model interpretation. Equally important is the "What is NOT on screen?" pass, where you preemptively exclude distracting elements like extra people, text overlays, logos, or unintended animals.

Timeboxing gives your prompt temporal structure. Specify both clip length and beat count: "6 seconds, 2 beats: reveal then reaction." This prevents the model from stretching or compressing action unpredictably. When naming elements, maintain rigorous consistency—if you call something "Boy A" or "Red umbrella" in one prompt, use identical identifiers in subsequent prompts to maintain continuity across your sequence.

Finally, respect the single-scene constraint. Most text-to-video models struggle with multiple locations in one prompt unless you explicitly mark transitions with clear directives like "CUT TO" or "MATCH CUT." When in doubt, break complex sequences into discrete, well-defined single-scene prompts.

Camera, Lens, and Motion Control

Camera language transforms flat descriptions into dimensional storytelling. The difference between amateur and professional AI video output often comes down to precise camera control—not just what you show, but how the camera reveals it to the audience.

Begin with camera position, specifying both height and distance from the subject. "Eye-level" creates intimacy and equality, "waist-high" adds subtle vulnerability, "aerial" provides context and scale, while "extreme close-up" builds intensity. These aren't just technical terms—they're emotional choices that shape viewer perception.

Shot type selection gives you framing control. A **wide establishing shot** orients the viewer in space, a **medium shot** balances subject and environment, a **close-up** prioritizes emotion, and an **extreme close-up (ECU)** magnifies micro-details like a single tear or trembling hand. Name the shot explicitly to lock framing and prevent the model from drifting between scales.

1

Define Position

Specify camera height and distance—eye-level, aerial, ECU—to lock perspective and emotional tone

2

Name the Shot

Use explicit shot types: wide establishing, medium, close-up to control framing precision

3

Motion Verbs

Precise motion language: dolly in, slow pan, orbit, handheld drift, not generic "camera moves"

4

Parallax Depth

Add foreground elements—window frame, leaves, fence—for cinematic depth during movement

Lens language adds cinematic texture. **Wide-angle distortion** exaggerates space and creates urgency, **telephoto compression** flattens depth and isolates subjects, while **shallow depth of field** blurs backgrounds to direct attention. These aren't camera settings—they're visual vocabularies that shape how viewers process information.

Motion verbs demand precision. Replace vague "camera moves" with specific directives: **dolly in** for smooth approach, **push-in** for forward momentum, **slow pan** for deliberate reveal, **tilt** for vertical exploration, **orbit** for 360-degree perspective, or **handheld drift** for documentary authenticity. Pair these with speed and smoothness descriptors—"slow, steady," "micro-jitter handheld," or "locked-off tripod"—to control motion texture.

Advanced techniques include focus behavior and parallax cues. Call out focus pulls explicitly: "rack focus from foreground raindrops to subject's eyes in 2 seconds." For camera moves, add foreground elements like window frames, leaves, or fences to create parallax depth—the hallmark of high-production-value cinematography. Finally, specify stabilization intent: "gimbal-smooth" for polished commercial work versus "documentary handheld" for raw authenticity.

Style, Genre, and Visual References

Visual style is your aesthetic fingerprint, but it requires surgical precision. The most common mistake in AI video prompting is style overload—throwing 10+ contradictory style tags at the model and expecting coherence. Instead, practice restraint with strategic specificity.

Limit yourself to 2–4 style tags maximum, and choose them deliberately. "Cinematic realism, indie drama, muted palette" creates a cohesive visual language. "Vintage film noir, cyberpunk neon, painterly impressionism" creates chaos. Each style tag competes for influence, so fewer, well-chosen tags yield stronger results.

Instead of name-dropping directors or cinematographers—which models interpret inconsistently—describe the qualities you want. Replace "shot like Wong Kar-wai" with "soft naturalism, quiet intimacy, saturated color isolation." Replace "Fincher aesthetic" with "cold fluorescent realism, geometric composition, desaturated shadows." This quality-based approach gives models clearer instruction.

Color palette callouts unify frames across generations. Choose 3–5 palette descriptors and repeat them consistently: "ashen gray, teal shadows, warm amber highlights." This constraint prevents the model from making arbitrary color choices that break visual continuity across your sequence.

Style Restraint Formula

Maximum 2–4 style tags to maintain coherence. More tags = visual chaos and conflicting aesthetics.

Quality Over Names

Describe visual qualities instead of name-dropping creators: "soft naturalism" not "like director X."

Color Palette Lock

Choose 3–5 palette words and repeat consistently: ashen gray, teal shadows, amber highlights.

Texture and medium specifications add tactile dimension. State whether you want "film grain" or "digital clean," "anamorphic bloom" or "spherical sharpness," "painterly brushstrokes" or "stop-motion tactility." These cues shape not just look but feel—the difference between polished commercial and gritty documentary.

Anchor time period with era-specific production design: "late-90s suburban, beige walls, CRT TV glow" immediately signals temporal context. The realism dial lets you control stylization level: "photoreal" for documentary authenticity, "stylized realism" for elevated drama, "dreamlike surreal" for memory sequences, or "graphic novel" for heightened expression.

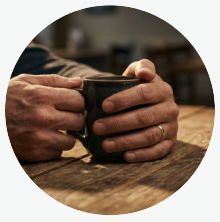
Lighting style functions as genre shorthand. "Noir chiaroscuro" evokes crime drama without explanation, "overcast natural" signals indie realism, "neon cyberpunk rim light" establishes futuristic tone. Finally, pair mood adjectives with visible proof: "melancholy (slumped posture, slow exhale, muted room)" grounds emotion in observable action rather than abstract feeling.

Subject, Performance, and Micro-Action

The difference between wooden AI video and compelling AI video lives in the granularity of performance direction. Vague emotional descriptions like "looks sad" give models nothing concrete to render. Micro-actions give them everything.

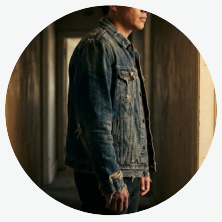
Replace "looks sad" with "eyes flick down, jaw tightens, forced half-smile fades." This level of specificity transforms abstract emotion into visible behavior the model can actually generate. Micro-actions are the secret language of performance direction in text-to-video prompting.

Create a body language list for every significant character moment, including 3–5 observable cues: posture (slouched, rigid, leaning), hands (clenched, fidgeting, open), gaze direction (averted, locked, scanning), pacing (hesitant, purposeful, erratic), and breath patterns (shallow, held, exhaled). This multipoint description gives the model multiple pathways to express the same emotional truth.



Micro-Actions

Replace vague emotions with specific behaviors: "eyes flick down, jaw tightens, forced smile fades"



Wardrobe as Story

Clothing details reveal character: "frayed denim, oversized hoodie, damp cuffs"

Props in Action

Give subjects tasks: folding paper, wiping fogged glass, tracing window condensation

Environment Reaction

Show world responding: wind catching hair, condensation forming, puddles rippling

Wardrobe functions as instant characterization. Don't just say "wearing casual clothes"—specify materials and wear patterns that tell stories: "frayed denim, oversized hoodie, damp cuffs" reveals socioeconomic context, emotional state, and recent activity in three descriptors. Clothing isn't decoration—it's narrative.

Face and emotion specificity requires calibration between "subtle" and "expressive," and naming the precise moment: "relief turning to dread" describes a transition, while "suppressed anger" suggests internal conflict. The more specific your emotional beat, the more nuanced the generated performance.

Interaction with props makes motion believable. Static subjects feel artificial; give them something to do—folding a paper crane, wiping fogged glass, tracing condensation on a window, adjusting a scarf against wind. Physical tasks ground performance in material reality.

For secondary characters, define their role and spatial relationship clearly: "background silhouettes only, out of focus, 15 feet behind main subject." This prevents the model from accidentally promoting extras to main characters or creating distracting competition for attention.

Finally, add environment reaction—how the world responds to the subject's presence. Wind catches hair, condensation forms on glass near warm breath, puddles ripple from footsteps, fabric lags behind turning movement. These physical interactions prove the subject exists in a coherent world rather than floating in an abstract void. Always keep identity descriptions generic, focusing on cinematic traits and performance rather than referencing real individuals.

Environment, Physics, and Continuity

Spatial Layout

Define foreground/midground/background elements so the model places objects consistently across frames

Weather Systems

Describe weather plus effects: "light rain, wet asphalt reflections, breath visible in cold air"

Named Light Sources

Specify key sources and direction: "streetlamp backlit, window side-lit, neon sign rim light"

Material Realism

Call out materials and behavior: "silk sheen, rust texture, dust motes in sunbeam"

Continuity Anchors

Repeat constants across clips: "same red scarf, same cracked tile, same graffiti tag"

Motion Physics

Describe gravity and inertia: "fabric lags behind turn, water drips downward"

Environmental coherence separates amateur from professional AI video. The physical world operates by consistent rules—gravity, light behavior, material properties—and your prompts must respect these laws to maintain believability.

Spatial layout establishes depth through explicit foreground/midground/background designation. "Foreground: chain-link fence, out of focus. Midground: subject walking left to right. Background: brick wall with graffiti, 20 feet behind." This three-plane description gives the model a depth structure to maintain across frames.

Weather isn't just a descriptor—it's a system with visible consequences. Don't just say "raining"—describe the ripple effects: "light rain, wet asphalt reflections doubling streetlights, breath visible in cold air, damp fabric clinging." Weather proves itself through interaction with the environment.

Named light sources with directional cues create motivated lighting. "Streetlamp backlit from above, window side-lit from left, neon sign rim light in teal" gives the model specific sources to render consistently. Motivated lighting feels intentional rather than arbitrary.

Material realism grounds your scene in tactile reality. Call out how materials behave: "silk sheen catching window light, rust texture on metal railing, dust motes visible in sunbeam." These surface qualities add richness and believability.

Continuity anchors are your multi-clip secret weapon. When generating a sequence, repeat exact constant elements across prompts: "same red scarf," "same cracked tile pattern," "same graffiti tag on wall." These repeated visual markers help viewers track continuity even when other elements shift.

Motion physics add credibility through observable cause and effect. Describe how objects respond to forces: "fabric lags behind as she turns, hair swept forward in wind, water drips downward leaving trails." Physics violations break immersion instantly—respect them explicitly.

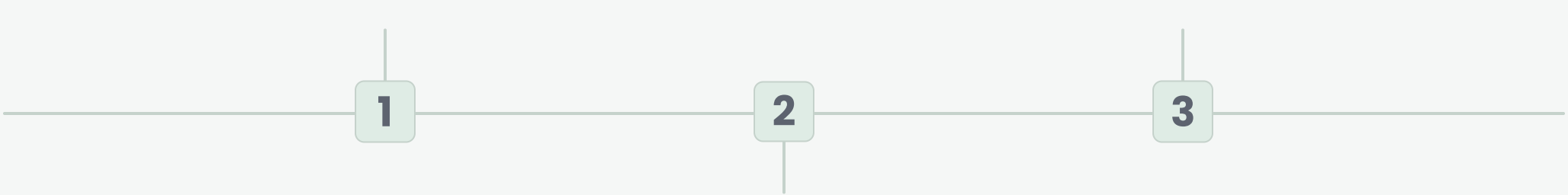
Scale cues prevent proportion drift across generations. Include reference objects in every prompt: "door frame 7 feet tall, car in background, standard dining chair." These anchors maintain consistent sizing, preventing the model from accidentally making subjects giant or miniature relative to their environment.

Finally, avoid impossible clutter. More props don't equal more realism—they equal "object soup" where everything competes for attention and nothing reads clearly. Prefer fewer, richly described items over many vague ones. "Single ceramic mug, white with hairline crack, half-full of cold coffee" beats "mug, papers, pens, books, phone, keys" in every scenario.

Editing, Timing, and Story Beats

Even when generating single clips, editorial thinking shapes pacing, rhythm, and narrative flow. Your prompt isn't just describing a moment—it's describing a beat within a larger sequence, and that beat has temporal structure.

Beat mapping gives your clip internal story progression. Write 2–4 beats with timestamp ranges: "0–2s reveal (door opens slowly), 2–5s approach (subject walks toward window), 5–6s reaction (pause, shoulders drop)." This temporal breakdown prevents the model from spreading action arbitrarily or front-loading everything into the first second.



Transition directives matter when planning multi-clip sequences. Use clear editorial language: "match cut" for visual continuity, "hard cut" for contrast, "fade" for time passage, "whip pan" for energetic connection. Only specify transitions when they're narratively motivated—most clips benefit from implicit cuts.

Loop-friendly endings enable seamless repetition if needed. Structure your clip so the final frame can lead back to the first: subject returns to starting pose, camera settles on establishing composition, motion completes its arc. Loopable clips give you flexibility in editing and allow for stylistic repetition effects.

Motivated camera ties camera movement to narrative cause. The camera shouldn't drift randomly—it should respond to story beats. "Push-in slowly as realization dawns on subject's face" or "pan right to follow subject's gaze toward window" makes camera an active storytelling participant rather than passive observer.

Maintain screen direction across clips to avoid spatial confusion. If your subject moves left-to-right in one clip, maintain that directional consistency in subsequent clips. Screen direction violations—where movement suddenly reverses without motivation—disorient viewers and break spatial continuity.

Insert shots punctuate emotion and provide editorial rhythm. Call them out explicitly when needed: "close-up insert: trembling hands gripping edge of table, 2 seconds." These brief moments of intense focus create visual variety and emphasize key details that wide shots might miss.

Rhythm adjectives shape temporal feel. "Slow burn" suggests gradual escalation, "staccato" implies quick cuts between moments, "lingering" indicates extended holds on emotional beats. Pair these rhythm descriptors with specific timing: "slow burn, 8-second shot with minimal movement."

Finally, consider silence versus sound cues. Even though most text-to-video models don't generate audio, descriptors like "silent, tense" or "ambient city noise implied" can shape visual pacing and performance intensity. Sound descriptors influence how the model interprets motion and expression, making them valuable even in audio-less generation.

Reliability Tricks and Failure-Proofing

Text-to-video models are powerful but temperamental. Even with perfect prompts, outputs sometimes drift, hallucinate, or misinterpret. These reliability techniques stack the odds in your favor, reducing failure rates and improving consistency across generations.

The constraint priority list separates must-haves from nice-to-haves. Structure your prompt with explicit hierarchy: "Must: fogged glass, two boys, dim hallway. Optional: distant basketball hoop, overhead fluorescent." This tells the model what to protect even if it has to sacrifice other elements. When processing power or attention runs short, the model knows what matters.

Negative prompts function as guardrails against common artifacts. Keep them short and specific: "No subtitles, no watermark, no extra limbs, no text overlays, no logos." Don't write negative novels—5–8 exclusions maximum. Too many negatives confuse the model as much as too few.

Priority List

Separate must-haves from optional elements explicitly in your prompt

Short Negatives

5–8 exclusions maximum: no subtitles, watermarks, extra limbs

Reduce Ambiguity

Replace "a thing" with "ceramic mug" or "folding chair"

Sanity Check

End with validation: "Only one person. No text. Slow dolly."

Reduce ambiguity nouns by replacing generic terms with specific ones. "A thing" becomes "ceramic mug," "furniture" becomes "folding chair," "structure" becomes "chain-link fence." Every vague noun is an invitation for the model to hallucinate something unintended. Specificity is reliability.

When continuity breaks across generations, simplify aggressively. Reduce to single focal subject, single action, single environment, single camera move. Complexity is the enemy of consistency. Once you achieve reliable simple output, layer complexity back incrementally while monitoring for drift.

Iterative tightening addresses persistent issues. If outputs consistently miss your intent, remove style flourishes and restate core constraints more plainly. Sometimes models get distracted by aesthetic descriptors and lose focus on fundamental structure. Stripping back to basics often reveals where the breakdown occurs.

Consistency tokens are reusable exact phrases that anchor elements across multiple generations. If you achieve a good result with "muted teal hallway" or "overcast window light," save those exact phrases and reuse them verbatim in subsequent prompts. Consistency tokens build a vocabulary of reliable descriptors.

Avoid conflicting styles unless you explicitly want hybrid aesthetics. Don't mix "cartoon" with "photoreal," "film noir" with "bright commercial," or "handheld documentary" with "locked-off symmetry" without clearly stating your intent. Style conflicts create visual mud where no aesthetic fully emerges.

The "sanity check" line is your final safety net. End every complex prompt with a short validation statement that restates critical constraints: "Only one person on screen. No text. Slow dolly in. 6 seconds." This redundancy catches common failures before they happen, giving the model one last chance to self-correct against typical artifacts.

Remember: reliability comes from clarity, hierarchy, and consistency. These aren't creative constraints—they're creative enablers that free you to focus on storytelling rather than troubleshooting. Master these techniques and you'll spend less time fighting the model and more time crafting compelling video narratives.