Rewriting Video: Text as Interface for Video Repurposing

SITONG WANG, Columbia University, USA
ANH TRUONG, Adobe Research, USA
LYDIA B. CHILTON, Columbia University, USA
DINGZEYU LI, Adobe Research, USA

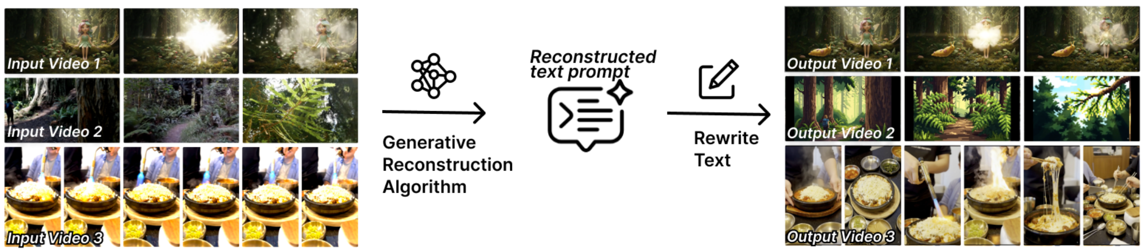

Fig. 1. Workflow of our text-as-interface approach for video repurposing. A generative reconstruction algorithm extracts an editable text representation from an input clip, which can then be modified via text rewriting to generate a new video output. See the complete list of use cases in Table 1. Here we highlighted (1) adding a new character; (2) changing the style to pixel art; and (3) diversifying the camera viewpoint.

Video is a powerful medium for communication and storytelling, yet repurposing existing footage remains challenging. Even simple edits often demand expertise, time, and careful planning, constraining how creators envision and shape their narratives. Recent advances in generative AI suggest a new paradigm: what if editing a video were as straightforward as rewriting text? To investigate this, we present a tech probe and a study on text-driven video repurposing. Our approach involves two technical contributions: (1) a generative reconstruction algorithm that reverse-engineers video into an editable text prompt, and (2) an interactive probe, Rewrite Kit, that allows creators to manipulate these prompts. A technical evaluation of the algorithm reveals a critical human-AI perceptual gap. A probe study with 12 creators surfaced novel use cases such as virtual reshooting, synthetic continuity, and aesthetic restyling. It also highlighted key tensions around coherence, control, and creative alignment in this new paradigm. Our work contributes empirical insights into the opportunities and challenges of text-driven video repurposing, offering design implications for future co-creative video tools.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Video repurposing, Text-driven video editing, Generative video models, Creative AI tools

1 INTRODUCTION

Video is one of the most powerful media for communication and storytelling, yet repurposing or adapting existing video content remains remarkably challenging. For professional and personal use alike, seemingly simple editing tasks demand significant technical expertise and time. Achieving a smooth transition may require deliberate camera movements during filming; capturing a scene from multiple angles often depends on a complex multi-camera setup. A single missed shot can be catastrophic to the intended narrative, and advanced techniques like reference-based editing

Authors' addresses: Sitong Wang, Columbia University, New York, NY, USA, sw3504@columbia.edu; Anh Truong, Adobe Research, New York, NY, USA, truong@adobe.com; Lydia B. Chilton, Columbia University, New York, NY, USA, chilton@cs.columbia.edu; Dingzeyu Li, Adobe Research, Seattle, WA, USA, dinli@adobe.com.

are largely inaccessible to everyday creators. These barriers fundamentally restrict how people can shape video into the narratives they envision.

Recent advances in AI—particularly in video understanding and text-to-video generation—offer a new paradigm for lowering these barriers. What if repurposing a video were as straightforward as editing text? This vision recasts video not as a fixed artifact, but as a medium with an exposed, editable "script." This script is not merely a transcript of spoken words; it encompasses the intricate layers that constitute a video: its visual compositions, rhythmic pacing, and overall aesthetic. This concept moves beyond conventional timeline manipulation or simple transcript-based editing toward a more holistic, semantic form of video authoring.

Unlike simply editing a transcript, which primarily focuses on the spoken word and its textual representation, we treat all modalities of the video as inherently editable components. This paradigm empowers creators to manipulate the video's essence with the fluidity of revising a written document. They could directly alter the visual flow, modify the temporal rhythm, or fine-tune other elements that shape the video's impact. The aspiration is to democratize the editing process by making all aspects of video manipulable through a script-like interface, unlocking new possibilities for creative expression and content adaptation.

To explore the potential and challenges of this text-driven approach, we developed a tech probe and conducted a probe study. Our work proceeds in three stages:

- We first developed a generative video reconstruction algorithm that iteratively reverse-engineers a source video into an editable text prompt. Our evaluations show the algorithm's accuracy converges after 3-6 iterations for a variety of video styles.
- Our analysis revealed a critical human-AI perception gap: while AI metrics prioritized frame-level fidelity (e.g., object accuracy), human viewers valued temporal-narrative coherence—qualities like pacing, vibe, and smooth transitions.
- Building on these insights, we designed and built an interactive probe system that exposes the editable prompts and integrates conversational assistance. The interface allows users to edit prompts directly and provide visual anchors (e.g., the first frame) to guide the generative process.

This work investigates the following research questions:

RQ1: How do creators envision using text-driven video repurposing tools, and what novel use cases are unlocked?

RQ2: What challenges and opportunities arise when video editing is mediated through text, and what are the design implications for future systems?

Through a probe study with 12 creators, we identified emerging practices such as video inpainting, B-roll variation, multi-angle synthesis from a single capture, and reference-guided adaptation of style and rhythm. These findings show how text can operate not just as an interface, but as a creative instrument—enabling creators to reshoot, remix, and restyle footage through language. Our work also reveals a central tension: while text-driven workflows can democratize video production by lowering technical barriers, they introduce new challenges in ensuring coherence, control, and alignment between human intent and generative output.

Together, these insights highlight the potential and current limitations of text-driven video repurposing, pointing toward future systems that support more accessible, expressive, and semantically aware forms of video creation. The contributions of this paper are:

- A generative reconstruction algorithm that enables videos to be deconstructed and reauthored through editable text prompts.

- The identification and analysis of a human-AI perception gap in video generation, where AI fidelity metrics diverge from human priorities of narrative and temporal coherence.
- An interactive probe, Rewrite Kit, demonstrating how text-based interfaces can support exploratory and iterative forms of video creation.
- Empirical insights from a probe study identifying novel practices, challenges, and design opportunities for text-driven video repurposing.

2 RELATED WORK

Our work is situated at the intersection of video manipulation, multimodal AI, and text-driven creative interfaces. We review these three areas to position our contribution.

2.1 Creative Manipulation of Videos

Video manipulation has long been central to creative production. Previous HCI systems for text-based editing [6, 9, 11, 18, 22] treated transcripts as semantic interfaces for navigating and rearranging footage. This approach made video editing more legible by turning it into something "writable;" yet its expressivity was largely limited to literal, selective operations on *speech-heavy content*. While such tools simplified tasks like selecting engaging clips [22] or inserting B-rolls [9], they offered little capacity to alter camera perspective, scene composition, or visual style—elements core to narrative transformation.

More recent AI systems have expanded this scope. Keyframe-guided inpainting methods [8] can synthesize motion and composition while preserving temporal coherence, and tools like Runway [19] allow users to restyle or extend footage through text prompts. However, these systems still tend to operate on a local level, treating video as a sequence of frames to be modified. This makes them powerful for substitution or restyling but less suited for enacting higher-level narrative or structural changes.

Our work builds on this trajectory but shifts the focus from localized editing to holistic, generative rewriting, where text is used to reimagine the video's narrative, pacing, and aesthetic structure. We explore how creators use language to reshoot, remix, and re-style existing footage.

2.2 Multimodal Generative Models

The technical foundation for our work lies in recent advances in multimodal generative models. In the image domain, models like CLIP [16] and principles like cycle consistency from CycleGAN [26] established robust connections between vision and language, laying the groundwork for unified systems that combine understanding and generation [25]. These ideas of alignment and consistency laid the groundwork for unified workflows that combine understanding and generation within a shared framework.

In the video domain, this progress is also accelerating. Large-scale models like Veo 3 [24] are demonstrating generalist temporal reasoning, while text-to-video generative systems such as Runway Gen-4 [17] and OpenAI Sora 2 [14] achieve impressive generation quality. Despite this progress, challenges in long-range continuity and narrative control persist.

Together, these developments reflect rapid progress in unified multimodal generative models. Crucially, however, rather than interactive creative use, so far the research focus remains largely model-centric, prioritizing accuracy and synthesis fidelity. Our work builds on these advances but shifts the focus toward human-AI co-creation, exploring how generative models can support text-driven video repurposing.

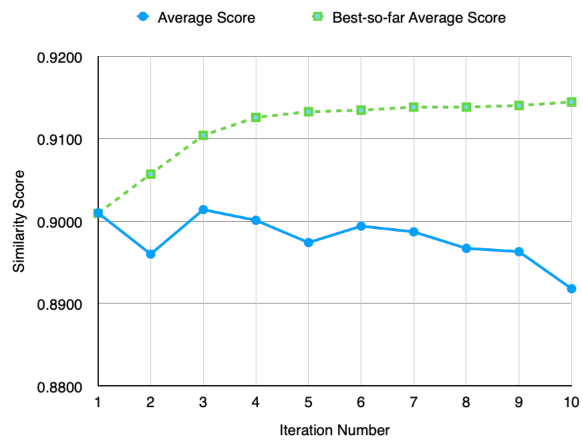

Fig. 2. The average similarity across iterations and the rationale for early stopping. The solid line shows the mean similarity score if we keep iterating uniformly, while the dashed line shows the "best-so-far" average, calculated as if each clip stops the first time it reaches its peak score. Performance gains typically saturate by 3-6 iterations. Further iterations often lower the per-iteration average due to "prompt drift," whereas the best-so-far curve remains flat, demonstrating the benefit of implementing per-clip early stopping.

2.3 Text-Driven Creative Interfaces

Our interface design builds on HCI research exploring natural language as a co-creative medium. In writing, systems like CoAuthor [12] have studied the collaborative dynamics between humans and language models. In visual art, tools like Promptify [2] and PromptCharm [23] support the iterative refinement of text-to-image prompts. This body of work shows that language can lower creative barriers but also shifts authorship toward a practice of prompting, steering, and interpretation.

Building on this foundation, HCI research has begun articulating design principles for prompt-based creation. Work such as design guidelines for prompting text-to-image models [13] and How to Prompt [5] frames prompting as a form of creative reasoning—one that involves articulating intent, interpreting model feedback, and negotiating meaning through language. These insights foreground the expressive and cognitive challenges of text-mediated creation, but prior work has largely centered on static text or image artifacts.

Our work extends this line of inquiry to video, where language functions as an editable substrate for repurposing and transformation. Rather than optimizing prompts for fidelity, we investigate how creators rewrite video—using language to author temporal, stylistic, and narrative change. This perspective situates text-driven video repurposing within broader efforts to design human-AI co-creative systems that center text as a medium of authorship.

3 GENERATIVE RECONSTRUCTION ALGORITHM

To enable text-driven video repurposing, we first proposed a method to deconstruct a source video into a faithful textual representation—an editable prompt that can accurately reproduce the original footage. We developed a generative reconstruction algorithm that achieves this through a closed-loop process of prompt generation, video synthesis, and comparative analysis.

3.1 Iterative Reconstruction Pipeline

Our pipeline reverse-engineers a video into a prompt through iterative refinement. Each cycle consists of generating a video from a candidate prompt, comparing it to the original, and using a vision-language model (VLM) to suggest improvements to the prompt for the next cycle.

3.1.1 Input and Initialization. The pipeline takes a video clip as input and uses a VLM (Gemini 2.5 Pro) to generate an initial descriptive prompt. The VLM is instructed to create a detailed, temporally-aware prompt suitable for a text-to-video model [7], based on an analysis of the video at 16 FPS. This output serves as the seed prompt for the first iteration.

3.1.2 Generation and Comparison. The candidate prompt is passed to a text-to-video model (Veo 3), conditioned on the first frame of the source video to maintain compositional consistency. The resulting synthetic video is then compared against the original. The same VLM (Gemini 2.5 Pro) analyzes both videos and generates a structured difference report, highlighting semantic gaps and providing a revised, improved prompt.

3.1.3 Iterative Refinement. The revised prompt from the comparison step becomes the input for the next cycle. This generation–comparison–refinement loop continues until a convergence criterion is met, such as when similarity scores plateau. In our experiments, performance typically peaked within 3–6 iterations (Figure 2). The final output is the prompt that yielded the highest similarity score during this process.

3.2 Automated Evaluation

We conducted an automated evaluation to measure the algorithm's convergence behavior and effectiveness.

3.2.1 Dataset and Metrics. We curated a dataset of 30 diverse 8-second video clips spanning genres such as vlogs, animations, cinematic scenes, and social media content. To quantify the alignment between a generated video and the original, we computed the frame-aligned cosine similarity between their CLIP (ViT-B/32) embeddings. This score serves as our primary metric for reconstruction accuracy.

3.2.2 Results. We ran the pipeline for 10 iterations on all 30 clips. The iterative process successfully improved upon the initial prompt in 80% of cases (24 of 30). The mean similarity score peaked at 0.9145, an average improvement of +0.0135 over the initial prompt's score of 0.9010.

As shown in Figure 2, performance improved rapidly in early iterations, with most clips reaching their peak similarity between iterations 3 and 6 (mean best iteration = 4.23). Beyond this point, gains diminished. Extended refinement often led to "prompt drift," where the model over-optimized for minor details at the expense of the core content. By iteration 10, the mean similarity had fallen to 0.8918, below the initial baseline. This suggests that early-stopping criteria are crucial for optimal performance. Clips with fast or complex motion (e.g., dance videos) proved more challenging, achieving lower peak scores and converging later.

3.3 Human Evaluation and the Perceptual Gap

To complement our automated metrics, we conducted a small-scale human evaluation to assess the perceptual quality of the reconstructions.

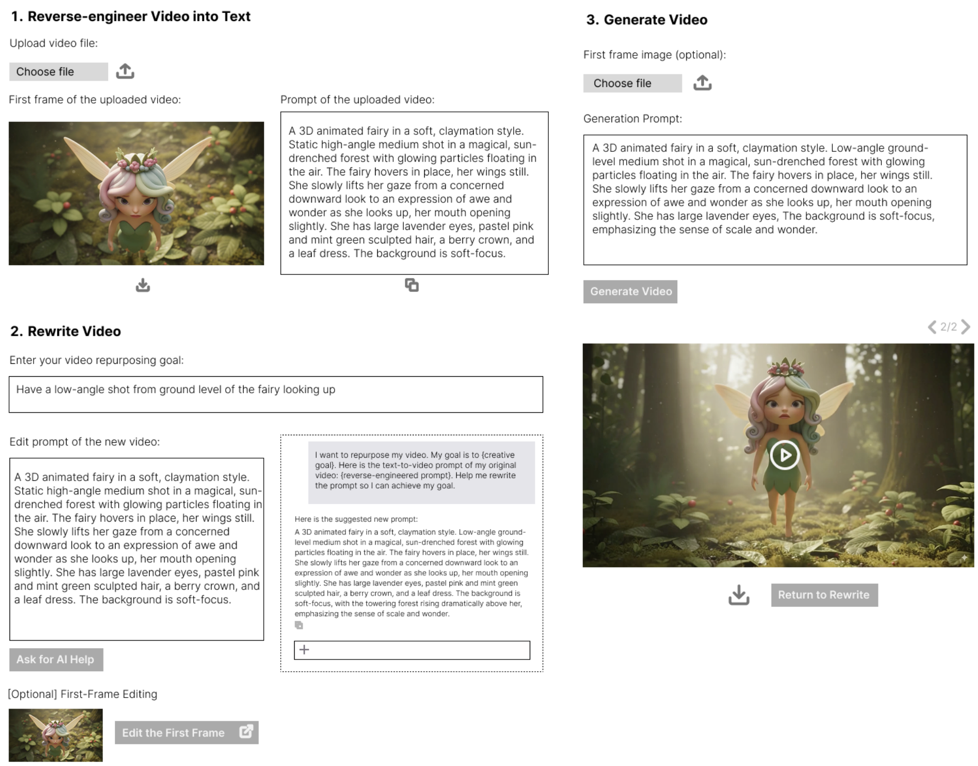

Fig. 3. The user interface of the Rewrite Kit technology probe. A creator's three-part workflow: reverse-engineer, rewrite, and generate.

3.3.1 Method. Two independent raters (Cohen's Kappa $\kappa = 0.82$) viewed all 30 pairs of original and best-reconstructed videos. They rated how well the reconstruction reproduced the original on a 1-7 scale and described the main differences. Raters were compensated $25 per hour.

3.3.2 Results: A Human-AI Perceptual Gap. The mean rating across all clips was 5.07 out of 7, with 96.7% of reconstructions judged as acceptable (≥4/7). However, the qualitative feedback revealed a critical *perceptual gap* between human and AI evaluation.

Our automated CLIP-based metric, like the VLM's comparison reports, emphasized frame-level visual fidelity: object accuracy, color, lighting, and composition. In contrast, human evaluators prioritized temporal and narrative qualities: the "vibe," pacing, rhythm, and realism of motion (e.g., choreography, character gait). Participants would rate a reconstruction lower due to a subtle difference in timing or movement, even when individual frames were visually accurate.

This finding suggests that while current AI models excel at seeing *what* is in a video, human perception hinges on *how* it unfolds over time. The essence of a video, from a human perspective, lies in its temporal flow and coherence. This highlights a key challenge for future work: developing models and interfaces that can reason about and manipulate the rhythmic and narrative qualities that make video a compelling medium.

4 DESIGN OF THE PROBE: REWRITE KIT

To investigate how creators engage with text as an interface for video repurposing, we developed Rewrite Kit, a web-based technology probe (Figure 3). Our design was guided not by the goal of creating a finished, optimal editing tool, but by the principles of a technology probe [1, 10]: to surface user strategies, interpretations, and tensions when using text to reimagine video.

4.1 Design Goals

Our design was shaped by an iterative process involving a research team with expertise in HCI, AI, and filmmaking. We established two primary design goals:

DG1 *Expose the "Script" of a Video*: The system needed to translate the implicit visual and temporal structure of a video into an explicit, editable textual artifact. This would provide a concrete object for creators to manipulate and reason about.

DG2 *Support Open-Ended, Exploratory Rewriting*: The interface should impose minimal workflow constraints, allowing creators to freely experiment with different ways of repurposing their footage. It needed to support both direct manipulation of the text and provide scaffolding for those who might need help articulating their goals.

4.2 The Rewrite Kit Interface and Workflow

The probe is organized around a simple three-part workflow: reverse-engineer, rewrite, and generate.

4.2.1 Reverse-engineering Video into a Textual "Script". Upon uploading a video, Rewrite Kit automatically runs our generative reconstruction algorithm. This pipeline reverse-engineers the video into an optimized text prompt that, along with the video's first frame, tries to faithfully reconstruct the original content. We run the algorithm for six iterations, at which point our technical evaluation showed converging accuracy. The resulting prompt is presented to the user in an editable text box next to a thumbnail of the video's first frame (Figure 3, Left), fulfilling DG1. This prompt becomes the primary medium for creative manipulation.

4.2.2 Rewriting the Script: Direct and AI-Assisted Prompting. To support a flexible rewriting process (DG2), the interface provides two modes of interaction. Users can directly edit the text prompt in the editor, treating it like any other piece of text. Alternatively, they can invoke a conversational AI assistant (powered by GPT-5) by clicking "Ask for AI Help." This opens a chat window where users can describe their creative goals in natural language. To scaffold this interaction, the probe provides a starter message: "I want to repurpose my video. My goal is to {creative goal}. Here is the text-to-video prompt of my original video: {prompt}. Help me rewrite the prompt. . . " Users can then easily copy the AI's suggestions back into the editor or continue the conversation to refine them.

For transformations involving significant stylistic or compositional changes, a revised text prompt alone may not be sufficient. Therefore, Rewrite Kit allows users to optionally provide a new first frame as a stronger visual anchor. Users can select "Edit First Frame," which uses an image generation model (Gemini 2.5 Flash Image) to create a new starting image based on their textual goals. The probe suggests a template for this request: "I want to repurpose my video. My goal is {creative goal}... Suggest an image-editing prompt to get the first frame of my new video." The user can then use the generated image in the final step.

4.2.3 Generating and Iterating on the New Video. Once the user has finalized their rewritten prompt and optional new first frame, they proceed to generation. This stage uses the Veo3 model to produce the repurposed video. The result is displayed in an embedded viewer, with a toggle that allows users to compare different generated versions and their corresponding prompts. A “Return to Rewrite” button enables a tight iterative loop, allowing users to go back, tweak the text or image, and regenerate as many times as they wish.

5 PROBE STUDY METHODS

We conducted a qualitative probe study to understand how creators would engage with a text-driven video repurposing workflow in practice.

5.1 Participants

We recruited 12 video creators (7 female, 5 male; age M=27.4, SD=3.1), each of whom creates or edits videos at least weekly. Our recruitment targeted a mix of user groups to gather diverse perspectives:

- **10 novice creators**, representing the primary target users of our tool, were recruited via university mailing lists and word of mouth. All novices had prior experience shooting and editing videos and were proficient with conversational AI tools. Half had previously experimented with text-to-video generation models. Novices were compensated $25 per hour.
- **2 expert creators** were recruited for their deep domain knowledge. Each had over 10 years of professional video editing experience and extensive familiarity with AI-based generation tools. One expert also had formal training in film production. Experts were compensated $50 per hour.

5.2 Procedure

Each participant attended a one-on-one session lasting approximately 45 minutes. Sessions were conducted remotely, and all interactions and dialogues were recorded with participant consent. The session was structured in three parts.

First, we introduced the concept of text-driven video repurposing and provided a brief demonstration of the Rewrite Kit probe’s functionalities. We explicitly framed the study’s goal not as an evaluation of usability or task performance, but as a way to elicit participants’ **explorations, reflections, and aspirations** when reimagining video through text.

Next, we invited participants to engage in an open-ended creative task. They were asked to bring one of their own video clips and brainstorm a real-world repurposing goal they wished to achieve. They then used Rewrite Kit freely to work toward this goal, with the aim of producing a new video they would consider “publishable.” Participants were encouraged to think aloud throughout this process. They were also permitted to use their preferred video editing software for post-processing if they felt it was necessary to finalize their vision.

Finally, each session concluded with a semi-structured interview. We asked participants to reflect on their experience, focusing on the challenges they faced, surprising outcomes, and the opportunities they envisioned for such a tool in their creative practice.

6 PROBE STUDY FINDINGS

We analyzed the study data using thematic analysis [3]. The first author conducted an initial coding of the transcripts and session recordings, from which the research team collaboratively developed the final themes. The findings are organized around our two research questions.

	User	User goal	Inputs	Desired transformation	Category
1	P1	Change the shooting angle of the dog running	Clip + ref. image	New camera angle	Re-shooting / perspective change
2	P2	Combine the concert stage and selfie	Two clips	Composited video	Remix / compositing
3	P3	Add dynamic camera motion to dance footage	Clip	Simulated moving camera	Re-shooting / motion synthesis
4	P4	Recreate the viral "historical selfie" clip for another figure	Ref. clip	Stylized reenactment	Re-worlding / generative narrative
5	P4	Remove a couple blocking the view in boat footage	Clip	Clean continuous shot	Remix / object removal
6	P5	Make a YouTube intro like a favorite channel	Ref. clip	Personalized intro	Re-styling / branding
7	P6	Recreate glitchy hoodie-swap transition	Ref. clip	Viral transition	Re-styling / transition
8	P7	Add a drone view of the national park	Clip	Synthetic aerial shot	Re-shooting / synthetic viewpoint
9	P8	Stylize vlog as pixel art	Clip + ref. image	Stylized render	Re-styling / aesthetic transfer
10	P9	Expand short food clip into rich vlog	Clip + ref. clip	Multi-angle, music-synced vlog	Re-styling / narrative enhancement
11	P10*	Add a yellow slug to the fairy animation	Clip + ref. image	New character inserted	Re-worlding / object addition
12	P10*	Change fairy shot to low angle from ground	Clip	Reframed spatial composition	Re-shooting / scene reframing
13	P10*	Change the camera angle of cat-riding-slug animation	Clip	Alternate camera view	Re-shooting / perspective variation
14	P11*	Generate a smooth transition between two clips	Two clips	Seamless temporal bridge	Remix / transition synthesis
15	P12	Extend the basketball shot to walking toward the hoop	Clip	AI-generated continuation	Remix / scene extension

Table 1. In the probe study, we found 15 potential use cases from 12 participants, spanning a wide diversity of video editing goals, inputs, and desired transformations.

Fig. 4. Use case 12: Camera angle change. Change fairy shot to low angle from ground.

6.1 RQ1: How do creators envision using text-driven video repurposing tools? What novel use cases are unlocked?

Participants used Rewrite Kit to repurpose their videos across a wide range of creative goals, treating text as a medium for **reshooting**, **remixing**, **restyling**, and **re-worlding** their footage. Across the fifteen distinct use cases we observed (Table 1), creators envisioned and enacted workflows that would traditionally require complex production setups or professional editing tools. These practices reveal how text-driven interfaces can unlock new forms of virtual filmmaking, synthetic continuity, aesthetic adaptation, and narrative world-building.

6.1.1 Text as a Virtual Camera: Re-shooting without Retakes. A striking pattern was creators using text to reshoot footage long after capture by changing the viewpoint, camera motion, or scene composition. Participants treated Rewrite Kit as a "virtual videographer," freeing them from the physical constraints of the original shoot.

For example, P1 turned a rear-facing clip of her running dog into a front-facing shot. She was surprised by the result:

> *"It's like I had a second camera in front of him - I didn't think I could manage to set it up in the real world."*

For her, text functioned as a directorial instruction: the words "from the front, wind in its fur" guided the system toward a shot she couldn't film alone. Similarly, P7 added synthetic drone views to their national park footage (as the use of drones was prohibited there). P3, a dancer, used text to add dynamic camera motion to a static shot, wanting to "boost the energy of my dance."

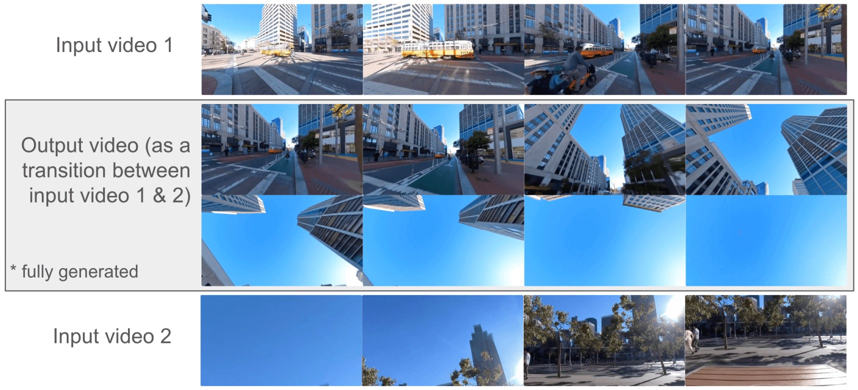

Fig. 5. Use case 14: Generate a smooth transition between two clips. Note that the middle two rows of transition are all 100% generated.

Expert P10, working with an animation (Figure 4), articulated how this could transform professional workflows:

> *"To really make an edit work, you want different angles... But it's such a hassle to get that coverage. This lets you expand any moment into a bunch of different angles you can actually cut between."*

She valued being able to describe a composition (e.g., "low-angle shot from ground level"), edit in text prompt, and generate variations quickly, and see the result match what she "pictured in my mind."

Across these cases, text became a virtual camera-level control surface, making complex cinematography achievable through natural language specification.

6.1.2 Text as Editorial Glue: Remixing and Bridging Footage. Another common practice involved using text to repair, combine, or connect footage. Here, creators treated Rewrite Kit as an AI editor that could blend disparate clips, clean up visual obstructions, or extend scenes to improve narrative continuity.

For instance, P2 combined a wide shot of a concert with a selfie video to create an imagined moment, describing it as "patching a memory I forgot to record." P4 used text for a restorative purpose: removing a couple who blocked the view in his speedboat footage. P12 extended a basketball shot to create a bridging scene for his vlog.

Expert P11 used the probe to generate seamless transitions between two otherwise disconnected clips (Figure 5), a task that typically requires meticulous planning during a shoot. He noted:

> *"Transition generation is cool. In traditional shooting, you have to plan... for a motion transition... Here, you can just generate."*

This family of practices points to what we call **synthetic continuity**: generating connective material that maintains narrative flow without manual keyframing or masking. With text as an interface, continuity became something creators could describe rather than manually construct.

6.1.3 Text as a Stylist: Restyling for Vibe. Many participants used Rewrite Kit to transform the look, tone, or rhythm of their videos, effectively adopting the role of an art director. Text, often supplemented by reference videos or images, served as a vehicle for aesthetic direction.

P5, creating a YouTube intro, used a reference from a creator she admired to "develop similar things for my own channel," a goal she found difficult to achieve before. P8 transformed her forest vlog into a pixel-art style (Figure 6),

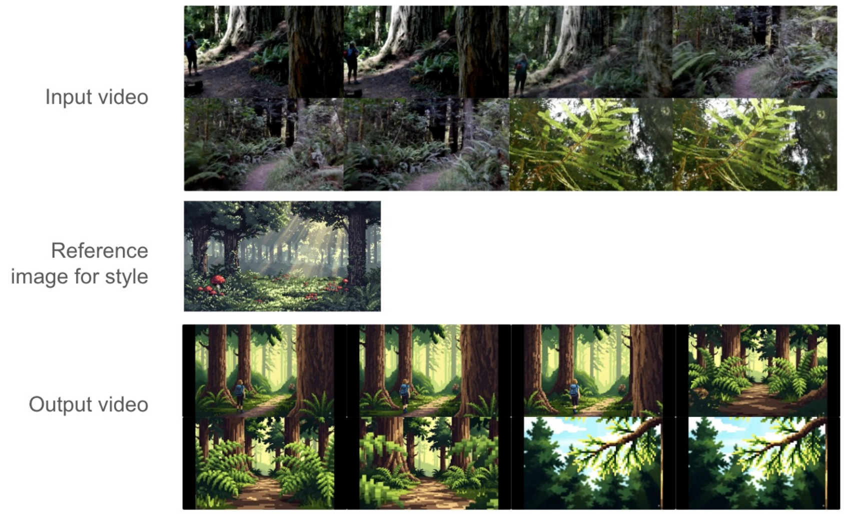

Fig. 6. Use case 9: Stylizing a vlog as pixel art. Guided by a single reference image, Rewrite Kit restyles the original photorealistic footage into a pixel art animation.

Fig. 7. Use case 10: Expanding a short food clip into a rich vlog. The input video, shot from a single static viewpoint, lacks narrative richness. Using a reference vlog as a guide, Rewrite Kit generates new viewpoints and synthesizes a more dynamic and engaging vlog.

while P9 used a reference vlog to expand a simple clip of melting cheese into a richly edited, multi-angle narrative (Figure 7).

A key insight from these sessions was that the reverse-engineered prompt from a reference clip served as a "scaffold for creativity." It provided novices with the necessary vocabulary to understand how a "vibe" was constructed, transforming an aesthetic they appreciated into a manipulable interface they could adapt for their own style.

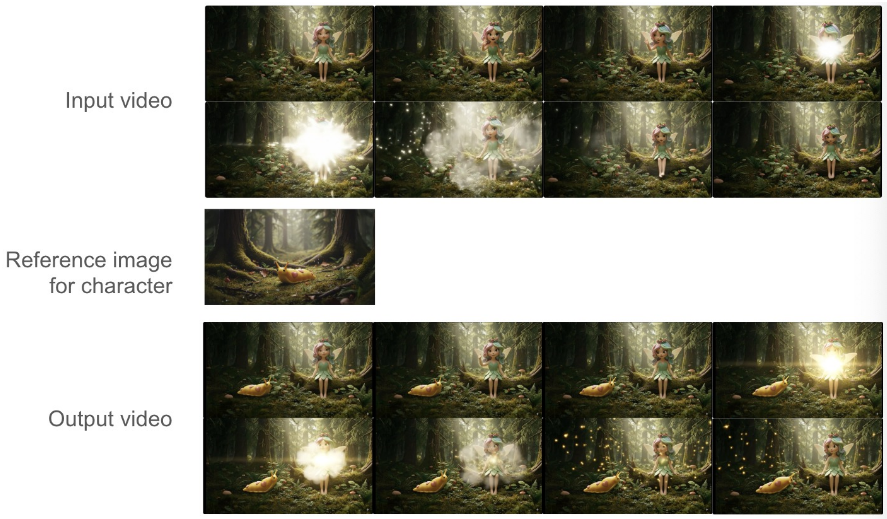

Fig. 8. Use case 11: Adding a yellow slug to an animation. For a participant wanting to insert a novel character from their asset library into an existing clip, Rewrite Kit provided a seamless workflow.

Fig. 9. Use case 4: Recreating a viral "historical selfie" with a new character. Inspired by a popular video of a historical figure taking a selfie, the user aimed to recreate the underlying concept with a different subject. This involved more than simple style transfer–Rewrite Kit adapted the original narrative's structure to fit this new character.

6.1.4 Text as a World-Builder: Re-writing the Scene's Reality. Finally, a few participants approached Rewrite Kit as a world-building engine, using text to rewrite *what a scene is about* rather than how it is shot. They inserted new characters, re-authored relationships, and reimagined worlds, transforming video editing into a mode of narrative redesign.

Expert P10 explored this by inserting a new character—a yellow slug—into her animated fairy clip (Figure 8). This act edited the scene's **narrative ontology**—the set of entities and events that exist within its world. P4 engaged in a similar practice, taking inspiration from a viral AI video of a historical "selfie" to restage the concept with a different figure (Figure 9). He reflected, "AI makes it feel like you can rewrite history visually."

In these cases, text enabled creators to treat videos as narrative spaces open to reinterpretation. The ability to modify who appears, what they do, or the world they inhabit suggests a powerful shift, transforming video from a record of captured reality into a writable, malleable universe.

6.2 RQ2: What challenges and opportunities arise when video editing is mediated through text? What are the implications for future design?

While text-driven editing proved powerful, it also surfaced deep tensions around coherence, control, and creative expression. These frictions were not simple usability errors; they were sites of tradeoffs where creators grappled with the boundaries between their intent and the model's interpretation. We identified four recurrent challenges that map the expressive and ethical landscape of this new paradigm.

6.2.1 The Tension of World Coherence: Keeping the Scene "True". A primary challenge was maintaining the internal logic of a scene after a generative change. Creators judged results not just on realism, but on whether new elements "belonged"—whether they inherited the scene's light, physics, and atmosphere.

P10, who added a slug to her fairy animation, wanted the slug to "sink into the moss—to feel a little more integrated." Her feedback framed coherence as a *relational property* between objects, not just visual plausibility. Similarly, P2 felt the lighting on her composited selfie was "a bit unnatural," and P1 noted that without a reference photo, the AI-generated dog felt like "a random dog," not her own. These moments highlight a new form of creative labor: **world-keeping**, or ensuring the rewritten world still "holds together."

Design Opportunity: Future systems should help creators reason about relational coherence. Instead of just placing objects, interfaces could allow users to define how new elements inherit world properties like illumination, physics, or causal logic, helping to sustain the scene's internal truth.

6.2.2 The Tension of Authenticity: The Real vs. Synthetic Boundary. Participants' tolerance for generative imperfection varied sharply depending on the source material. In animated or fantastical contexts, approximation was acceptable. But for footage of real people or cherished memories, even minor flaws felt intrusive. P10 articulated this asymmetry:

> *"I'm much more flexible when the modifications are on an animated... video... But if it were a clip of me or my kids, I'd have a higher bar for preservation."*

The most common fault line was the nuance of movement. P3 found the system "not excellent at reproducing choreography," while others noted that a dog's jog appeared "slightly mechanical" or a speedboat's up-and-down motion was lost. These micro-failures broke the illusion of reality, reminding participants of the generative layer underneath. This boundary was not only aesthetic but also ethical, as creators implicitly balanced when an edit crossed from representation into fabrication.

Design Opportunity: Instead of chasing flawless realism, systems could introduce an "authenticity gradient"—a control that lets creators choose what kind of truth a scene should preserve: visual, emotional, or narrative. Making this gradient explicit would encourage reflection on where a work lies between documentary fidelity and imaginative synthesis, helping creators stay intentional about the truths their creations convey.

6.2.3 The Tension of Translation: Aligning Creative Voice with a Reference. Participants consistently used reference videos not to copy them, but as expressive catalysts for developing their own creative voice. They sought to extract and adapt specific qualities—a style, rhythm, or structure—for their own content. The reverse-engineered prompt often served as a "creative schema," giving them the vocabulary and structure to deconstruct and repurpose an aesthetic.

However, a tension arose when the text failed to capture the essence of a reference. P9, working on a food vlog, felt the results were close but wanted to "adjust the beats to match the music like the reference did," noting that non-visual rhythm was the heart of its appeal. This highlights that repurposing is a complex act of translation and negotiation—deciding what to keep from a reference, what to transform, and how to infuse it with a personal voice. For P4, a viral "George Washington selfie" video became inspiration for restaging the idea with a different historical figure—a form of visual parody that blended homage with reinterpretation. He collaborated closely with the conversational AI to co-develop a script, ensuring that the new narrative remained coherent while preserving the meme's core structure. These examples illustrate that text-driven repurposing is not a simple pipeline from description to output, but an ongoing negotiation of creative alignment—deciding what to keep, what to transform, and what to invent.

Design Opportunity: Systems could support this translation process by helping users deconstruct a reference into its core components (e.g., pacing, color palette, camera motion). By making these underlying principles explicit and editable, tools can help creators borrow structural ideas while developing a distinct creative voice.

6.2.4 The Tension of Modality: When Words Are Not Enough. Despite the power of text, participants inevitably hit expressive ceilings where language felt insufficient. Cinematic qualities like gesture, rhythm, and composition often resisted linear, verbal description.

8 of 12 participants used the optional first-frame editing feature, finding that providing a visual anchor gave them more reliable control. Others expressed a desire for more visual and direct manipulation. P8 wanted a storyboard to preview keyframes before the "heavy video generation process," while P10 wished she could "directly annotate on the frame and tell the model what happens next." When collaborating with the AI, many participants naturally began constructing multimodal prompts, combining images and text to convey their intent more precisely.

Design Opportunity: The "perfect prompt" is a myth. A good prompt is often a combination of text, image, and cues in other modalities. Future systems should treat text as one channel within a broader multimodal conversation. Interfaces that allow creators to fluidly interleave describing, showing, and marking will better support the embodied, iterative nature of creative thought.

7 DISCUSSION

Our study illustrates how text-driven video repurposing reconfigures creative work, shifting it from frame-level manipulation toward high-level semantic creation with generative models. While the preceding sections identified empirical patterns, here we discuss the broader implications of our findings for creative agency, authorship, and the design of future co-creation tools.

7.1 From Direct Manipulation to Semantic Generation

Text-driven interfaces transform video repurposing from a process of manual control into semantic writing and generation. Rather than manipulating pixels on a timeline, creators articulate their intent through language and iteratively steer the system toward a desired outcome. This reframes authorship as a conversational act centered on expressing, interpreting, and refining intent with an AI partner. This model does not replace the HCI tradition of direct manipulation [20], but complements it with a paradigm grounded in semantic expression and shared agency. Future co-creation tools should therefore prioritize interpretability—making system reasoning visible and editable—over simply providing precise control of low-level parameters.

7.2 The Perceptual Gap as a Design Material

Our evaluation revealed a persistent **human-AI perceptual gap**: while models optimize for frame-level fidelity, human judgment hinges on temporal rhythm, emotional tone, and narrative coherence. Rather than viewing this gap as a technical flaw to be eliminated, it can be treated as a design material for reflection. Future interfaces could expose the system's interpretation of "similarity," allowing creators to consciously decide when to prioritize visual accuracy versus narrative flow, depending on their own taste and preference. Such transparency would elevate generative tools from opaque assistants to reflective collaborators, supporting more intentional control over the temporal and affective meaning of a video.

7.3 World-Keeping: A New Form of Creative Labor

Generative transformations introduced a distinct form of creative work we term **world-keeping**: the labor of sustaining the internal logic, physics, and atmosphere of the source material while rewriting its content. A successful edit was not defined by realism alone but by whether it "belonged"—whether its light, motion, and mood cohered with the original scene. While this world-keeping will improve as generative models become more capable, we believe it is critical to develop interfaces that make relational coherence an explicit, editable property, enabling creators to specify what dimensions of reality (visual, emotional, causal) they wish to preserve, instead of offloading all the decisions to the opaque models. In this sense, world-keeping marks a new locus of creative authorship: the craft of maintaining continuity in an inherently fluid, generative medium.

7.4 Learning Through Re-creation with Reference Media

Creators consistently used reference clips not to copy, but to learn how a particular "vibe" is constructed. By reverse-engineering prompts from examples, they gained a scaffold for understanding the constituent elements of a style—pacing, camera work, color—which they then re-contextualized in their own work. This process treats exemplars as analyzable, decomposable models of style rather than as fixed templates. Future systems could make this deconstruction explicit, breaking references into editable layers (e.g., rhythm, tone, motion dynamics) to better support creative learning and reinterpretation over one-click style transfer.

7.5 Beyond Text: Toward Multimodal Authoring

Although text afforded powerful high-level control, participants repeatedly turned to visual and temporal cues—first frames, examples, or imagined storyboards—when language reached its expressive limits. This highlights the need for multimodal authoring environments where text, image, gesture, and timing operate as complementary channels of intent. As underlying models become increasingly multimodal, the design challenge shifts to orchestrating these inputs fluidly, allowing creators to describe with words, demonstrate with visuals, and mark through temporal cues within a single iterative loop.

8 LIMITATIONS AND FUTURE WORK

First, our probe study focused on short video clips (typically under 10 seconds), reflecting the current capabilities of text-to-video models. Future work should explore how these text-driven practices scale to longer narratives, which may require new techniques for maintaining global continuity across decomposed scenes.

Second, this work is situated within a rapidly evolving technological landscape. We did not evaluate our system with the most recent models (e.g., Sora 2 [14]) since it came out only days before the paper submission deadline. While we expect our core design insights around control, intent alignment, and coherence to hold, future studies should revisit these ideas with next-generation architectures.

Finally, our prototype prioritized a text-first workflow. Integrating other emerging modalities, such as annotation-based or sketch-based video manipulation, could yield richer, more embodied forms of control while preserving the conceptual clarity of text-driven interaction.

Ethical Considerations. Text-driven repurposing amplifies critical ethical concerns. The ease of rewriting footage blurs boundaries of authorship and can be used to copy a creator's stylistic identity [15] or manipulate real-world imagery in deceptive ways [21]. Addressing these risks is both a technical challenge and a design imperative. Future systems should be built on a foundation of responsibility, embedding safeguards such as provenance tracking [4], style attribution, and consent indicators to foster a culture of transparent and ethical creative practice.

9 CONCLUSION

In this paper, we investigated a new paradigm for video creation: repurposing footage as fluidly as rewriting text. Through our generative reconstruction algorithm and the *Rewrite Kit* probe, we demonstrate that exposing a video's underlying textual "script" empowers creators to reshape narratives through language, enabling practices like virtual reshooting, restyling, and remixing that move beyond timeline constraints. Our findings reveal both the promise and the friction of this approach: while text enables powerful, intent-driven control, it also surfaces fundamental challenges around coherence, multimodal expression, and creative alignment. Ultimately, this work reframes video repurposing as an act of expressive writing, pointing toward a future of co-creative systems that make visual storytelling more fluid, interpretable, and fundamentally more accessible.

10 GENERATIVE AI USAGE DISCLOSURE

Generative AI tools were used solely to refine the authors' own writing for clarity and grammar. No generative AI tools were used for project ideation, study design, coding, data analysis, or result interpretation.

REFERENCES

[1] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI interprets the probes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1077–1086.

[2] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[4] Coalition for Content Provenance and Authenticity (C2PA) 2025. *C2PA Specification, Version 2.2*. Coalition for Content Provenance and Authenticity (C2PA). https://spec.c2pa.org/specifications/specifications/2.2/index.html Accessed: 7 October 2025.

[5] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390* (2022).

[6] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

[7] Google AI. 2025. Generate videos with Veo 3 in Gemini API: prompt guide. https://ai.google.dev/gemini-api/docs/video?example=dialogue#prompt-guide. Last updated 2025-10-06 (UTC).

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Chenlin Meng, Omer Bar-Tal, Shuangrui Ding, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2025. Keyframe-Guided Creative Video Inpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13009–13020.

[9] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J Mysore. 2019. B-script: Transcript-based b-roll video editing with recommendations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[10] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[11] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. 2017. Computational video editing for dialogue-driven scenes. *ACM Trans. Graph.* 36, 4 (2017), 130–1.

[12] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[13] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–23.

[14] OpenAI. 2025. Sora: Creating Video from Text. https://openai.com/index/sora/. Accessed: 2025-10-07.

[15] Julien Porquet, Sitong Wang, and Lydia B Chilton. 2025. Copying style, Extracting value: Illustrators' Perception of AI Style Transfer and its Impact on Creative Labor. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[17] Runway Research. 2025. Introducing Runway Gen-4. https://runwayml.com/research/introducing-runway-gen-4. Accessed: 2025-10-07.

[18] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-based tools for editing audio stories. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 113–122.

[19] Runway ML. n.d.. Creating with Aleph. https://help.runwayml.com/hc/en-us/articles/43176400374419-Creating-with-Aleph. Accessed: 2025-10-07.

[20] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 08 (1983), 57–69.

[21] The New York Times. 2025. A.I. Videos Have Never Been Better. Can You Tell What's Real? *The New York Times* (June 29 2025). https://www.nytimes.com/2025/06/29/technology/ai-videos-real-or-fake.html Accessed: 2025-10-07.

[22] Sitong Wang, Zheng Ning, Anh Truong, Mira Dontcheva, Dingzeyu Li, and Lydia B Chilton. 2024. PodReels: Human-AI Co-Creation of Video Podcast Teasers. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 958–974.

[23] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[24] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. 2025. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328* (2025).

[25] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. 2025. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295* (2025).

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.