

Multimodal Agents Are Becoming Creative Teams

Video, voice, documents, slides, websites, and visual-to-code workflows are converging around a shared brief. The opportunity is to carry creative decisions through an entire project—and make the next revision easier.

Mykola Lobanov, Founder & Chief Product and Technology Officer
October 1, 2026 · 2649 words

A product launch rarely needs just one piece of content. It needs a clear story, a landing page, a sales presentation, a product video, supporting documents, and often versions for different audiences and languages. Each deliverable depends on decisions made elsewhere: the approved product description, the visual direction, the evidence behind a claim, the action the audience should take.

That coordination is becoming part of what AI agents can do.

In its July 2026 introduction of ChatGPT Work, OpenAI described a workflow that carries customer research into a campaign brief, creates marketing assets, and adapts them for different markets. In September, ElevenLabs expanded its MCP connection to make speech, transcription, dubbing, music, sound effects, images, and video accessible from an assistant, with generated assets available in its creative workspace for further editing. These are documented product capabilities; the quality of a complete campaign still depends on the inputs, execution, and review.^{[1](#), [2](#)}

What makes a system a creative team

The direction is significant. Creative work is moving toward systems that can participate in several stages of production while retaining the context that connects them.

The phrase “creative team” describes that division of work. A system can help research a subject, shape a narrative, generate media, assemble deliverables, and check the results. One coordinating agent with specialized tools may perform those roles. Some workflows may benefit from several agents with separate responsibilities. The useful measure is how well the work stays connected.

Multimodality and agency contribute different things. Multimodality lets a model interpret or generate different media. Agency lets a system select tools, take actions, inspect results, and continue toward an objective. Documents, presentations, and websites add another requirement: they are structured artifacts with headings, objects, layouts, links, or executable behavior. Producing them often involves code or application tools alongside the model.

This combination makes a broader creative workflow possible. A recording can inform a document. The document can supply the narrative for a deck. A design can guide a website. A reviewer can point to a visual problem, and the agent can use that feedback to revise the underlying artifact.

Video: an editable cut, not a finished clip

Video makes the change particularly visible. In May 2026, Google introduced a Flow Agent that can support brainstorming, creation, editing, variations, batch changes, and asset organization. Its Gemini Omni integration brought conversational video editing into the same environment. That connects more of the work around a shot: developing an idea, exploring alternatives, and maintaining a usable collection of assets.³

For a product marketer, an effective video workflow could begin with real product footage, an approved script, a shot list, and visual references. The agent could help develop alternative openings, prepare supporting scenes, and assemble a reviewable cut. The editor would judge whether the pacing makes sense and whether the product is represented accurately.

The editable timeline matters. ElevenLabs' Studio Agent can place clips, generate voiceovers, find voices, and coordinate sound effects within its timeline. Its documentation describes a handoff in which a person can take manual control, make changes, and return the work to the agent. That is a concrete example of human and agent contributions sharing the same working artifact.⁴

Control also has limits. Google's August 2026 Gemini Omni model card identifies remaining challenges with consistency across edits, complex motion, and accurate text rendering. For branded work, this makes exact logos, interface labels, prices, and other essential details sensible candidates for controlled overlays or verified source footage. A visually convincing scene still needs to communicate the right facts.⁵

Voice, as the product and as the way to direct it

Voice participates in this workflow in two directions: it is a medium to produce and an interface for directing production.

On the production side, agents can help move from recordings to transcripts, scripts, narration, and localized audio. ElevenLabs describes its Dubbing v2 model as conditioning on the original performance to carry qualities such as pacing and delivery into the translated speech. That is the vendor's stated capability; whether a particular dub preserves meaning and works for its audience requires listening and language review.⁶

A useful voice workflow therefore retains an approved script, pronunciation guidance, the selected voice, and a record of permission to use it. Review should cover meaning, emphasis, timing, and whether the narration matches what is visible. A translated sentence may need different editing or screen time even when the translation itself is correct.

On the input side, Google's May 2026 Stitch update describes live design collaboration through text or voice, with existing code and design files providing context. Spoken direction can become part of an iterative visual conversation: a person describes what needs to change while looking at the work. The system's ability to connect that instruction to the relevant object or layout is what makes the interaction useful.⁷

Documents carry the evidence everything else depends on

Documents supply much of the substance that these experiences depend on. Research notes, interview transcripts, product specifications, and spreadsheets often contain the evidence behind a campaign. Agents can help turn that material into a brief, proposal, report, or customer guide while preserving references to the underlying sources.

Claude's file-creation documentation, for example, supports generating Word documents, PowerPoint presentations, spreadsheets, and PDFs through its computing environment. That gives an agent a route from source material to files people can download and continue working on.⁸

The production standard should extend beyond creating a valid file. A useful document needs accurate claims, a coherent argument, appropriate structure, readable tables, and a layout that survives export. Teams should check both the content and the rendered pages. An extracted text check can confirm that a paragraph exists while missing a table that has split awkwardly across pages.

Documents also make a practical place to establish the shared narrative. An approved brief can record the audience, product capabilities, supporting evidence, terminology, and desired action. Every downstream asset can then inherit those decisions. This reduces the number of places where a generated interpretation can quietly become a new product claim.

Slides compress the story without losing the argument

Slides introduce a different editorial challenge: compressing a narrative while preserving the reasoning that makes it persuasive.

Microsoft documents a Copilot workflow in PowerPoint that lets users refine an outline before generating slides, then continue editing the presentation. It also recommends starting with an organization's template so that the generated deck can use its theme and existing layouts. The source document's structure helps determine how the content becomes slides.⁹

For a sales deck, that suggests a sensible sequence: agree on the audience's problem, establish the argument, choose the evidence, and then develop the visual presentation. A slide should have a purpose in that argument. Attractive individual slides can still produce a confusing deck when they repeat claims, skip necessary explanation, or introduce inconsistent numbers.

Editability is another acceptance criterion. When the team needs to revise text, charts, and diagrams, those elements should remain accessible as working objects wherever practical. A deck made entirely from flattened images creates extra work for the next person who needs to change a number or translate a label.

Research illustrates how much this matters. PPT-Eval evaluates computer-use agents on 120 PowerPoint tasks across 12 files, covering creation and editing. Its published results report a 45% success rate for the tested Claude 4.5 Opus agent, alongside a 57% average partial score. Those figures apply to that benchmark and configuration; they do not measure every current presentation product. They do show why "the file was created" is an incomplete definition of success.¹⁰

Websites have to work, not just look right

Websites bring the creative workflow into an interactive environment. The copy, visual assets, and page structure must work together with navigation, forms, responsive layouts, and sometimes a backend.

Figma Make provides one example of that convergence. Its documentation describes generating interactive apps from prompts and attached designs, editing the code and preview, adding a backend, and publishing a website. The working preview gives teams something they can explore and revise while the implementation develops.¹¹

For a launch page, this can connect previously separate activities: adapt the approved narrative into page copy, apply the design system, place media, implement the call to action, and review the result in a browser. Acceptance should

follow the intended experience. Does the form reach the right destination? Does the mobile layout preserve the content hierarchy? Can someone use the page with a keyboard? Are the visible product claims consistent with the brief?

The implementation requirements grow with the project. A campaign page and a customer portal may share a visual style, but a portal also needs explicit decisions about identity, permissions, stored data, and operational ownership. Those requirements belong in the brief and validation process.

From a design to running code, and back

Visual-to-code workflows make the connection between design and implementation more direct. A screenshot can communicate composition, spacing, color, and hierarchy. A structured design file can additionally expose components, variables, and layout information. The repository contributes another layer: the components and conventions the product already uses.

Figma's documented Codex workflow retrieves design context for code generation and can bring a running interface back into editable Figma frames. Teams can revise that design and then carry the changes into the application. This creates a useful cycle between visual exploration and implementation.^{[12](#)}

The distinction between appearance and behavior remains essential. A static image does not specify what happens after a failed form submission, how a table behaves with hundreds of rows, or which components should be reused. A strong implementation brief pairs the visual reference with those requirements. The agent should then inspect the running interface and test the interactions that matter.

The Figma2Code research benchmark reinforces this point. Its authors evaluated ten models using design images and metadata and reported continuing limitations in responsive layouts and code maintainability, even when visual fidelity was stronger. That is evidence about the models and tasks studied, and a useful reason to evaluate generated interfaces beyond their first screenshot.^{[13](#)}

Six handoffs, and what a person still reviews

Across these six areas, the practical handoffs look like this:

Work	Connected workflow	What the team should review
Video	Brief, footage, and references into scenes and an editable cut	Narrative, continuity, timing, and product accuracy
Voice	Recording or script into narration and localized audio	Meaning, pronunciation, delivery, and synchronization
Documents	Source material into a structured, editable file	Evidence, argument, citations, and page layout
Slides	Approved narrative and data into a presentation	Story, chart accuracy, readability, and editability
Websites	Brief, design, and assets into a working site	User actions, responsive behavior, accessibility, and data handling
Visual-to-code	Visual references and design context into implementation	Component reuse, layout fidelity, interaction states, and maintainability

One brief, every deliverable: a launch worked through

The larger opportunity appears when these workflows share a source of truth.

Consider an illustrative SaaS launch. The team has customer interview recordings, product specifications, approved screenshots, and brand guidelines. It needs a product brief, an eight-slide sales deck, a landing page, and a 30-second video with localized narration. This is a proposed workflow, rather than a claim about a measured customer result.

1. Establish the source material. Extract themes from the interviews, connect proposed claims to evidence, and identify questions the product owner needs to resolve. Keep the original sources available throughout production.
2. Agree on the story. Approve the audience, core message, visual direction, product facts, and call to action. Resolve those shared decisions before generating a large collection of assets.
3. Produce the deliverables. Give each workflow the same approved brief, then adapt its expression to the format. A deck needs an argument; a landing page needs a clear path to action; a video needs a sequence that works in time.
4. Inspect the actual outputs. Read the rendered document, review the deck, play the video with sound, and use the website. Combine automated checks with human judgment about the message and experience.

5. Handle changes through their dependencies. If a feature name changes, identify the document passages, slides, page sections, captions, and spoken lines that depend on it. Revise and review those items while preserving unrelated approved work.

That fifth step is a demanding test of the “creative team” idea. It requires the system to remember where a decision was used and to make a targeted correction. The value can persist after the first draft, through the revisions that accompany real projects.

The supporting architecture can be straightforward. Keep a versioned brief, an asset collection, and a record of each deliverable’s sources and approval status. Use tools that preserve editable working files. Give each action appropriate access and distinguish drafting from publishing. Add separate specialist agents when a clear task boundary or parallel workload justifies the coordination they introduce.

MCP, the Model Context Protocol, can provide a common way for assistants to reach external tools and context. The ElevenLabs and Figma integrations show how that can connect media generation and design workflows. Teams still need to define what the assistant may change and who decides when the result is ready. A connection supplies access; the workflow supplies purpose and accountability.^{[2](#), [12](#)}

Cost per approved deliverable, and who decides

The economics should be measured at the point where the work becomes usable. A useful operating metric is cost per approved deliverable: generation, tool usage, orchestration, human review, and revision costs divided by the number of accepted outputs. Track the elapsed time from brief to approved package as well. Cheap drafts can create expensive review queues if the team generates more alternatives than it can meaningfully assess.

This favors early decisions about the story and progressive investment in production. Review an outline before generating a deck. Approve the script and storyboard before producing final video. Use draft media where it helps test an idea, and spend on higher quality output once the direction is sound. For revisions, measure whether the requested change succeeded while previously approved content remained intact.

Creative leadership remains central to that process. Someone has to decide which customer insight matters, which claim deserves emphasis, which visual direction fits the brand, and which version is ready for an audience. Agents can make those decisions easier to explore by producing concrete alternatives and carrying chosen directions into multiple formats.

For teams adopting multimodal agents, a repeatable workflow is a practical starting point: one launch package, one proposal process, or one localization pipeline. Give it shared sources, editable outputs, clear review criteria, and a way to propagate changes. The most useful creative system will be the one that helps the team make a good decision once—and apply it consistently wherever the work takes shape.

Notes

Research checked September 23, 2026. Product capabilities above are grounded in official documentation and announcements; availability can depend on plan, platform, and region. Benchmark findings describe their published test settings. Workflow recommendations and the launch scenario are editorial analysis.

1. OpenAI — ChatGPT is now a partner for your most ambitious work, July 9, 2026.
2. ElevenLabs — Introducing voice, music, image, and video generation in the ElevenLabs MCP, September 14, 2026; updated September 20, 2026.
3. Google — New agents, mobile apps and Gemini Omni for Google Flow and Google Flow Music, May 19, 2026.
4. ElevenLabs — Introducing Studio Agents, May 7, 2026; updated September 16, 2026.
5. Google DeepMind — Gemini Omni Flash model card, updated August 27, 2026.
6. ElevenLabs — Introducing Dubbing v2, May 28, 2026; updated September 6, 2026.
7. Google — New ways to design in real time with Stitch, May 19, 2026.
8. Anthropic — Create and edit files with Claude, current help documentation.
9. Microsoft — Create a new presentation with Copilot in PowerPoint, current support documentation.
10. Gandhi et al. — PPT-Eval: A Benchmark for Computer-Use Agents on PowerPoint Tasks, 2026.
11. Figma — Explore Figma Make, current product documentation.
12. Figma — Building Frontend UIs with Codex and Figma, February 26, 2026.
13. Gui et al. — Figma2Code: Automating Multimodal Design to Code in the Wild, April 15, 2026.