Digital "Double Hatters": Augmenting Audiovisual Creative Work with a Generative Text-to-Video Workflow

Vivian Liu
Columbia University
vivian@cs.columbia.edu

Tao Long
Columbia University
long@cs.columbia.edu

Jenny Ma
Columbia University
jm5676@columbia.edu

Nathan Raw
Rochester Institute of Technology
nathanrawdata@gmail.com

Jiaxin Yang
Columbia University
jy3318@columbia.edu

Claudia Tang
Columbia University
ct3008@columbia.edu

Lulu Wang
Columbia University
yw4240@columbia.edu

Yumo Yang
Columbia University
yy3204@columbia.edu

Lydia Chilton
Columbia University
chilton@cs.columbia.edu

Abstract

Generative AI is transforming work for creative professionals. Text-to-image and text-to-video tools provide alternative methods for visual content creation, circumventing the traditional workflows creative professionals have developed. This transformation raises the pressing question of whether these new generative workflows will augment or supplant creative work. To investigate this question, we introduce Generative Disco, a text-to-video system for music visualization, and use it as a technology probe for creative professionals who freelance audiovisual work. In a mixed-methods study (n=12), we observe that professionals found Generative Disco easy to use, highly expressive, and capable of supporting many professional use cases. Its generative workflow enabled professionals to become digital "double hatters", expanding their creative range into adjacent domains. We conclude on how creatives can benefit from this new means of skill mobility.

Keywords: generative AI workflows, creative work, text-to-image, large language models, music visualization, social media, creativity support

Generative AI has changed the landscape of creative work. A group of people who are directly impacted are freelance creatives and independent artists. Previously, clients would seek these creative professionals in talent marketplaces for jobs involving design, video, and content creation work. Creative professionals would be hired based off of their skills, clients would hand professionals concepts to fulfill, and they would work together towards a desired outcome.

Generative AI has disrupted this traditional exchange of creative work by presenting people with powerful new tools for content creation. Text-to-image and text-to-video models allow users to write prompts to generate images and video that can express a near infinite range of visual concepts (Brooks et al., 2024; Dayma et al., 2021). These workflows are generally simple and declarative, and they present alternatives to the steep skill verticals associated with traditional tools like Photoshop or Premiere. These "low floor, high ceilings" advantages have led to the rapid adoption of tools such as DALL-E and Midjourney and the emergence of AI-generated content on social media.

Another driving force behind the adoption of generative AI is the neverending demand for visual content. To visually represent the different things we create and disseminate on media platforms, we need art. When a song is released, it has to be accompanied by album art, music videos, and visualizers. When news is published or books are released, they have to be paired with news illustrations and book covers (Liu et al., 2022). Generative methods provide people with efficient and personalizable means to do so but raise the question if such methods will augment or displace creative work (Eloundou et al., 2023).

We explore how these new generative workflows can impact creative professionals who freelance audiovisual work. To focus our scope, we study how music and video professionals engage with a text-to-video workflow for music visualization. We choose music visualization, because it is a richly multimodal task fusing music and video that is representative of the complex projects creative professionals are tasked with. Music visualization is an activity that is core to social media — people create content around viral songs and sounds, and a niche of AI-generated content consisting of generated music videos and dance animations has

URI: https://hdl.handle.net/10125/109164
978-0-9981331-8-8
(CC BY-NC-ND 4.0)

emerged on platforms like TikTok.

For this task, we introduce Generative Disco, a text-to-video workflow we built for music visualization. The tool presents an interactive pipeline incorporating large language model assistance, text-to-image generation, and text-to-video generation to help users take a music file as input and produce an animated music video as output. The tool introduces design patterns called transitions and holds that users can interactively apply to scaffold the construction of text-to-video narratives. An example music visualization from the tool is shown in Fig. 3.

We conducted a first-use user study with video and music professionals (n=12) who tried Generative Disco and compared and contrasted its generative workflow with their own expert ones. We present qualitative findings showing that generative workflows enable audiovisual professionals to be digital "double hatters" by allowing them to augment their skillsets and expand their creative range into adjacent domains. Music professionals were empowered to visually represent their work and concretely convey the vibes and visuals carried within music. Video professionals could extend their range into visual spaces that were previously outside of their usual technical capabilities, such as stylized animation and morphing. We additionally report use cases professionals described of how they could use AI to overcome time and resource constraints and also quantify the perceived usefulness and ease of use of the generative workflow. We conclude by discussing how digital double hatting can enable horizontal expansion across domains and the implications of such skill mobility for creatives.

1. Background

1.1. Generative AI in Creative Workflows

Machine learning advances in modeling multimodal knowledge have led to meteoric improvements in generative technologies. These advancements have translated into products such as Stable Diffusion, Midjourney, DALL-E, and Sora. Given a prompt, these image and video generative models are capable of producing a limitless amount of visual content that expresses a vast range of concepts and aesthetics. While prompt engineering for image generation still remains a trial and error process, many of the initial challenges in producing high quality outcomes have been solved by advancements in design guidelines, crowdsourced galleries, and developments in control (Liu & Chilton, 2022; Zhang & Agrawala, 2023).

As generative models have matured, researchers have applied generative workflows to real-world problems that creative professionals face. Systems centered around creative writing, news angle exploration, and science communication have shown that large language models can be powerful ideation support tools for narrative tasks (Gero et al., 2021). Prior work has also shown that large language models can be chained with text-to-image models to streamline the exploration of generative design outcomes (Liu et al., 2022). Image generation pipelines have found use in multiple design tasks ranging from visual metaphor generation to storyboarding (Wang et al., 2023, 2024). Generative tools can be valuable for brainstorming, addressing design fixation, generating assets, and getting publishable outcomes. Industry standard tools such as Photoshop have also begun to ship generative features.

While text and image models have been more broadly studied and deployed, video models are on the horizon. Sora and ImagenVideo have shown that generative AI can produce high-resolution, cinematic video clips (Brooks et al., 2024; Ho et al., 2022). However, the lower accessibility of advanced models so far has led to less exploration of text-to-video workflows and their downstream impacts on creative work.

1.2. AI Impacts on the Creative Workforce

Creative work is often done by freelancers and artists who take on high-skill content creation work such as design and video editing. These are independent, self-employed workforces that often work under tight timeframes for variable compensation and without the support structure of a team. Studies of freelancers have shown that they are over two times more likely to use generative AI in their workflows than non-freelancers (Upwork, 2023). When creative professionals connect to clients through gig work platforms, they often have to operate within a reputation economy (e.g. five star system) and are rated after jobs for their skills, delivery, and communication (Yoganarasimhan, 2013). Such evaluations can add pressure and strain worker autonomy. Creative professionals on platforms such as Upwork also have the option to offer full-service packages (e.g. creating an explainer video or an animated logo), which can put often put creative professionals in situations where they handle the end-to-end delivery of a creative product alone.

Many of the industries and occupational tasks traditionally open to independent work have been suggested to have high exposure to generative AI. For example, Eloundou et al. (2023) conduct an analysis studying exposure across O*NET occupational skills

and find that programming and writing are the skills most positively correlated with exposure. Upwork (2023) empirically corroborates this, identifying research, translation, brainstorming, writing, and coding as points where freelancers turn to generative AI. A longitudinal deployment of a generative writing support tool found that technical experts do benefit from the efficiency gains of an AI tool beyond a novelty effect (Long et al., 2024). In a large-scale deployment of AI assistance for customer support agents, Brynjolfsson et al., 2023 found that AI can improve productivity and even the skill curve between workers by surfacing the tacit knowledge of the most productive workers to help the less experienced ones.

1.3. AI Content Creation for Social Media

Creative work often targets publication on social media. An estimated one quarter of freelance work is targeted for social media platforms such as Youtube, TikTok, and Instagram (Upwork, 2023). Shortform video is a predominant format on these platforms, and its creation process can often be a high-production effort to create involving scripts, music, voiceover, footage, and effects. In a study of Youtuber use of generative AI in videos, Lyu et al., 2024 found that creators were utilizing generative AI at each touchpoint of their creative process: ideating new video topics, drafting scripts, generating assets, upscaling existing content, and automating voiceovers. Brüns and Meißner, 2024 also found that generative AI has disrupted the marketing content creation process for brands and influencers and has been attenuating how consumers perceive brand authenticity on social media.

These factors have led to AI-generated content emerging as a new form of content on social media. As of June 2024, there are 79.9 million posts related to AI Art on TikTok and 15 million posts tagged with #aiart on Instagram. Users can now create personalized content with models as a form of participatory entertainment: people apply AI filters to place themselves as characters in alternative settings (image-to-image), generate music videos and dance animations to express music (text-to-video, video-to-video), and generate images to signpost messages (text-to-image) (Lyu et al., 2024). A survey of content creators (n=7000) found that 52% of content creators either extensively or occasionally use AI tools and that 22% plan for future adoption (Artlist, 2024). These findings motivate the need for studies focused on how multimodal AI capabilities (such as its capacity for audiovisual understanding and generation) can boost content creation on social media, which is often driven by trending music and visual stories.

2. System: Generative Disco

To explore a multimodal AI workflow for creative professionals engaged in content creation, we built Generative Disco, a text-to-video tool for music visualization that is illustrated in Fig. 1. It takes music in as input and generates music visualization in the form of shortform video as output. Its guiding design principles are to help users express music by 1) supporting simple prompt-based interactions, 2) structuring the creation of a text-to-video narrative, 3) helping users brainstorm visuals to interpolate into video, and 4) generating videos that have open-ended artistic possibilities.

2.1. System Walkthrough

2.1.1. Input Music and Interactive Segmentation. The system begins by taking a music file input and representing it within the interface as a waveform (Fig. 1-A). Users interactively choose how they would like to segment the music by adding intervals and editing the time boundaries delineating intervals. The waveform illustrates the music's structure and helps highlight volume dynamics, lyric boundaries, and other musical elements.

2.1.2. LLM-Based Prompt Ideation. To generate video for a music segment, users have to define prompts for images to START and END intervals on (Fig. 1-C). To streamline the prompting process we provide a BRAINSTORMING AREA (Fig. 1-D) which incorporates large language model assistance. Users provide goals for specific intervals (Fig. 1-B) and GPT-4 returns relevant visual subjects. For example, a user could type in *"dancing at the disco"* and receive subject suggestions such as *"Robot DJs, neon dance floor"* or *"Groovy dancers, colorful disco ball"*. Below the subject suggestions, Generative Disco also provides a palette of style keywords that can add aesthetic specificity to the prompt. These keywords were sampled from a superset of 100 style keywords that was sourced from prior work analyzing composition and style keywords for AI-generated art prompts (Liu et al., 2022). In Fig. 1-E, the selection of *"glitch"* and *"slow motion"* assembles the prompt: (*"Robot DJs, neon dance floor, glitch, slow motion"*).

2.1.3. Text-to-Image Visual Brainstorming. Generating a prompt generates a set of images from Stable Diffusion. Generations that users liked could be dragged to the interval START and END images for each interval to keyframe the text-to-video

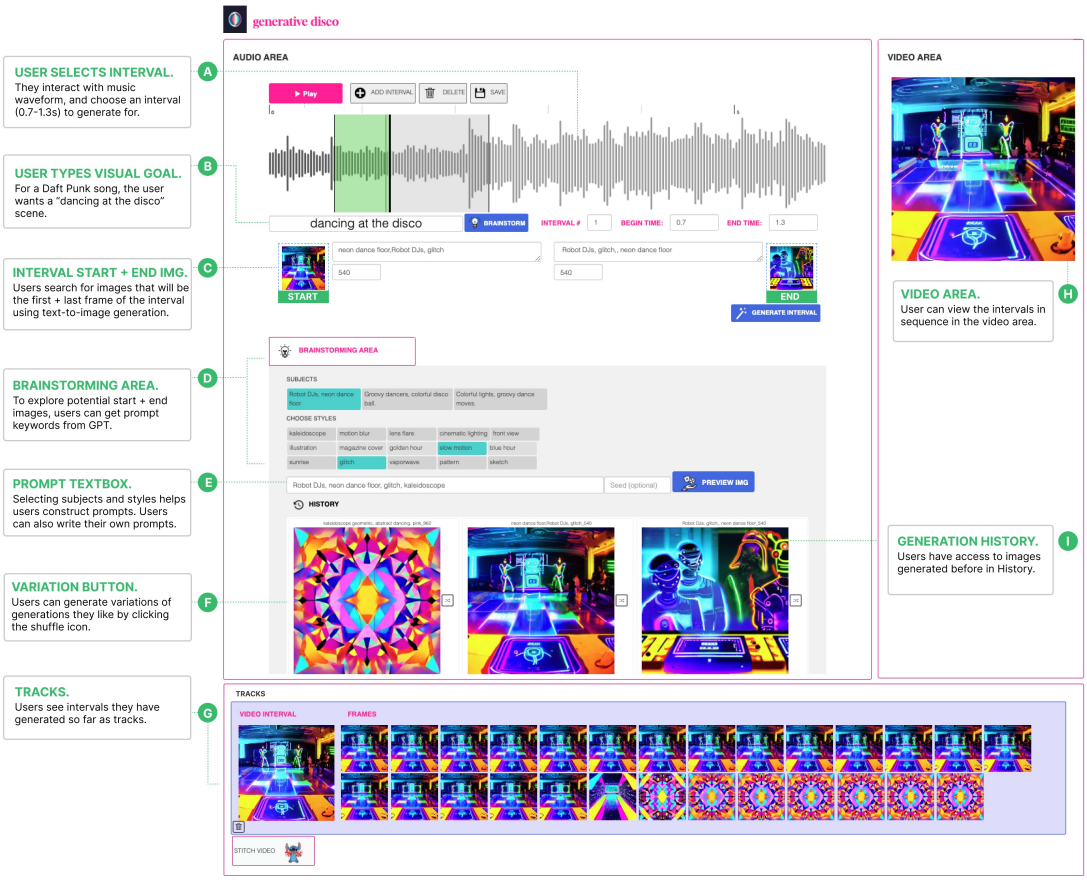

Figure 1. Generative Disco system design. Users begin by interacting with the waveform to create intervals within the music (A). To find prompts that will define the start and end of intervals (C), users can brainstorm prompts using suggestions from GPT-4 (B, D) and explore text-to-image generations (E, I). Results users like can be dragged and dropped into the start and end areas (C), after which a text-to-video interval can be generated. These show in the Tracks (G) and can be stitched into a video placed in the Video Area (H).

generation. Users could also create variations of generations (Fig. 1-F), which would regenerate images with automatically modified parameters (e.g. prompt with shuffled phrases, different seed hyperparameters). Generative Disco's visual brainstorming features allow users to work on filling out the music at a high-level and prototype different visual narratives.

2.1.4. Generating Text-to-Video Intervals. Once a user chooses a pair of START and END images, a music visualization clip can be generated by interpolating between them to the intensity and beat of the music. The interpolation is implemented by interpolating between the text embeddings of their associated prompts and noise latents. The audioreactivity is implemented by taking a beat-based analysis (harmonic percussive source separation) of the input music. Both capabilities are enabled by the open-source repository Stable Diffusion Videos ("Stable Diffusion Videos", 2022). Frames were collected together at 24 fps.

2.1.5. Output Video. Generated intervals (Fig. 1-G) are stitched together using the 'STITCH VIDEO' button. The final output video is placed at the top right of the interface in the VIDEO AREA (Fig. 1-H).

2.1.6. Text-to-Video Design Patterns. A challenge unique to text-to-video is that when generated videos are not conditioned on base images or video, they lack the intuitive physics and temporal constraints we expect from real video. In Generative Disco, we choose to guide the generation with only text prompts and not with initial image or video prompts. This way, the text-to-video outputs would not constrained by what is expected from reality or existing media and have the same open-sky possibilities allowed in art.

We provide two design patterns, transitions and holds, to help structure text-to-video narratives. Generated video has to have the right amount of visual interest: it should not be as still as a slideshow but it should also not be a video of nonstop change. **Holds** are a way to help focus the video on specific shots, highlighting subjects of interest. In contrast, **transitions** drive visual change, moving the narrative from subject to subject and shifting the color palette and style aesthetic to reflect the energy and emotions latent in music.

Generative Disco supported these design patterns by surfacing images suitable for implementing them. Holds could be implemented by parameterizing start and end frames of intervals with images that were highly similar (e.g. variations of one another). Transitions could be implemented by using the subject and style suggestion features to find pairs of images that had larger semantic and visual distances between them. Users could interactively apply a combination of holds and transitions to create a visual cadence that is acceptable to the human eye.

3. Evaluation

We conducted a mixed methods study, where 12 video and music experts were brought in to field test Generative Disco and compare its generative workflow against their expertise. This evaluation centered around the following research questions. **RQ1)** To what extent do freelancers find Generative Disco useful and usable? **RQ2)** How does Generative Disco impact the skillsets of creative professionals–does such a tool augment their skillset or supplant it? **RQ3)** What is the expressive range of Generative Disco as a generative framework: what does it succeed at visualizing and what are its failure modes?

Recruitment. Our participants were recruited from 1) Upwork, a platform for freelancers, where we reached out to creatives with professional video or music experience 2) independent musicians from a local computer music organization. Participants were paid $40 per hour for their time, and the study was conducted for 2 hours. Twelve people (6 male, 4 female, 2 non-binary) participated. Their average age was 26.5 (min=21, max=41). The experimental IRB protocol was approved by the institution.

Participant Backgrounds. Participants were first interviewed about their creative expertise and traditional workflows for video editing. They were also asked about their exposure to generative AI. Seven participants had exposure to generative AI tools such as ChatGPT and had used it professionally. Participants described using it to generate royalties agreements with clients (P12), write routines that mixed and mastered their original music (P10), and write video scripts for their clients (P3). P8 and P3 had both posted their own music online with text-to-image generations as cover art.

Research Protocol. After a brief interview about their professional experience, an experimenter explained concepts behind Generative Disco through a slide deck that gave them a primer about text-to-image generation, prompts, and hyperparameters (seeds). Afterwards, an experimenter demonstrated how to brainstorm prompts, generate images, and choose images to start and end intervals on the interface. Participants were then given a training task to playtest the design patterns of transitions and holds by generating them for a provided song. This helped them familiarize themselves with the kinds of outputs the system could generate.

Participant backgrounds are described in Table 1. Prior to the experiment, participants had selected a song of their choice and sent a short 10-15 second music clip. The experimental task was to generate music visualization for this clip. After completing the open-ended experimental task, participants filled out a post-study questionnaire and were interviewed about their experience. All text-to-video intervals and associated prompt pairs were also collected for analysis.

3.1. Quantitative Feedback

To answer RQ1, we first report quantitative metrics on Generative Disco's usefulness and ease of use.

3.1.1. Creativity Support Index Metrics Generative Disco performed well in terms of Creativity Support Index (CSI) metrics. These participant responses (all on a 7-point Likert scale), are visualized in the middle subplot within Figure 2. All 12 participants rated the system a 6 or 7 for enjoyment (median:7). Ten of 12 participants gave positive feedback (positive defined as $\geq$ 5 out of 7) that the results were worth their effort (median: 6.5). Eight of 12 participants agreed that they could sufficiently explore a number of outcomes without tedious interaction (median: 6). Expressiveness was similarly generally positive (median: 5.5).

There was a slight split in opinion on control *("I had control over the intervals and the video I was generating"*, median: 5) Nine of 12 participants rated it 5 or higher, but the remaining three (P1, P2, P9) found that they had lower ability to control the system, a problem that characterizes many generative workflows. (For example, if a user prompted for a person standing, they didn't have pixel-level control over how that person

stood.) There was a similar divergence of opinion for Ability (*"I generated videos I would have otherwise not been able to create."*). Nine of 12 participants were in agreement with this statement about Ability (median: 7). The remainder thought that creating such work was within their abilities but that Generative Disco could speed up the process.

3.1.2. NASA-TLX Metrics The majority of participants recorded very positive responses for system performance (median: 6). The vast majority also did not find the system to be frustrating, temporally demanding (median:2), mentally demanding (median: 2), or effort-intensive (median: 3). Almost every participant (11/12) responded that their frustration during the task was low (low defined as ≤ 3 out of 7, median: 2).

3.1.3. Workflow-Specific Questions The majority (7 of 12) rated the system positively for how closely it allowed them to engage with the music (median: 5). Nine of 12 rated the system positively for audiovisual alignment (*"The system helped me come up with visuals that aligned with the music."*, median: 5). Eleven of 12 participants were positive about the helpfulness of GPT-4 subject brainstorming (median: 5.5). Nine of 12 participants were positive that the style keyword brainstorming area was helpful (median: 5.5). When asked if Generative Disco would be a useful addition to their current video / music workflow, 9 of 12 participants responded positively for agreement (median: 5.5).

3.2. Qualitative Findings

To contextualize the quantitative feedback, we describe use cases reported by freelancers (RQ1).

3.2.1. AI for Overcoming Time and Resource Constraints Participants described how Generative Disco could help in situations were they are under time and resource constraints. P1 had made over 100 lyric videos for clients in their seven years of experience. They described how Generative Disco could solve a friction point (searching for stock footage) while also providing content that was uniquely music-aware.

> *"Looking for footage to use is very time consuming. It's probably my least favorite part of the process for my workflow... I've been using it [stock footage] all my life, but I don't even have access to amazing stock footage websites or anything, because sometimes it can get really expensive... I would say with this [Generative Disco], you gotta learn it, it's a whole new way of working... If I was doing it without its help, it would be a lot of me cutting to the music... but it wouldn't look as flawless the way that it does [here] with the transitions on the beat. " -P1*

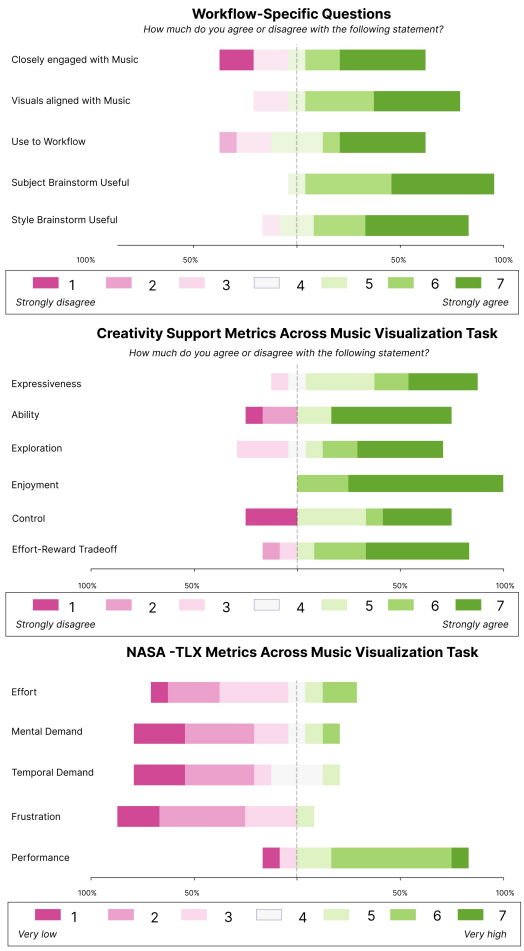

Figure 2. Participant responses on NASA-TLX, creativity support, and workflow-specific questions.

P5 similarly expressed that Generative Disco could greatly benefit their workflows as a footage generator.

> *"I couldn't "reach" that [transition] with actual footage from [stock footage site] or [stock footage site]. I would have to shoot it myself with my camera and hire a real actor... It would cost me my time, my energy."* - P5

Professionals mentioned that the generative capacity of AI tools can be especially helpful when they have to provide a long stream of content, as in livestreams, live production, or VJ loops (P4). *"A lot of content*

Table 1. Table of participant details: background (primary area first), engagement with video work, years of experience (of primary area), exposure to visual generative AI, and genre of music for task.

ID	Background	Video Freq	Yrs Exp	AI-Art Freq	Genre
P1	Video Pro, Lyric Videos	Daily	7	Never	Metalcore
P2	Video Pro, VJ	Daily	14	Never	Original Composition
P3	Video Pro	Daily	3	Weekly	Pop
P4	Video Pro, live production, VJ	Weekly	15	Weekly	Funk Rock
P5	Video Pro, Sound Design	Daily	5	Never	Alternative Indie
P6	Music Expert	Yearly	4	Yearly	Acoustic
P7	Music Expert, Classical + Digital	Monthly	5	Never	Hard Rock / Remix
P8	Music Expert, Acoustics + Production	Weekly	8	Monthly	Original Composition
P9	Music Expert, Video Pro	Yearly	10	Monthly	Dance / Electronic
P10	Video Pro, Music Videos	Monthly	10	Weekly	Locked Groove
P11	Video Pro	Daily	6	Weekly	Afrobeats / Pop
P12	Music Expert	Yearly	20	Never	Original Vocals / Rock

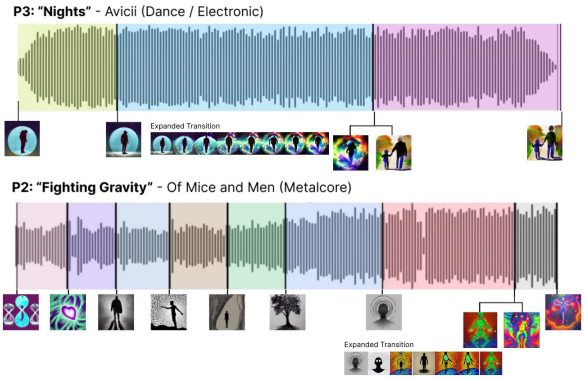

Figure 3. Examples of how participants segmented their music and keyframed their text-to-video music visualization. Two transitions are expanded for illustration. Forks between images indicate jump cuts.

for DJs doesn't need to be real clean. It can be busy. You're providing content for a half hour, so having stuff that you don't have to recycle, if you could have really long clips and premade stuff–that could benefit [VJs]." P4, P5, and P10 all described how Generative Disco outputs could be used as a mixing layer to merge and layer with real footage or as the background of a video. P10 additionally described that Generative Disco could help them test out color corrections, color boards, and scales, which to them *"tended to be the hardest part of making or doing anything with videos"*.

3.2.2. Skill Augmentation in Visual Work (RQ2) Creative professionals who specialized in video work described how Generative Disco gave them access to techniques and content that is often out of their reach.

For example, P3 described how Generative Disco made stylized animation more accessible to them when they created a music visualization for "Nights" by Avicii that is pictured in Fig. 3. To the beginning lyrics, *"Someday you'll leave this world behind..."*, P3 generated a hold of a shot of an astronaut in space. At the next lyric phrase *("so live a life you will remember")*, they applied a color transition, saturating the blue moon background into a rainbow one and expressing the lyrics with fullness and color. In the last lyric phrase, *"My father told me when I was young"*, they generated another hold on a father and son holding hands in a watercolor style. P3 commented that the stylized animation effect would have been outside of their technical expertise and required the enlistment of an animator.

P2 and P10 also described how Generative Disco enabled them to use other visual techniques that were new to them such as morphing. P2 generated a subject transition that morphed a human silhouette into a tree (Fig. 4-Row 5), finding this continuous and "dream-like" transition appealing and suitable for the atmospheric intro of their song. Another participant created a visual transition representing the passage of time by morphing a vibrant bouquet of flowers into decayed one (Fig. 4-Row 3).

Even for advanced freelancers, there can be visual techniques such as morphing and stylization that are gated by technical know-how. These examples illustrate how Generative Disco could expand the visual repertoire of a freelancer and help them move into adjacent visual spaces like animation.

3.2.3. Enabling Double Hatting: Music and Video (RQ2) Likewise, we found that Generative Disco could boost music experts into the visual space. P8,

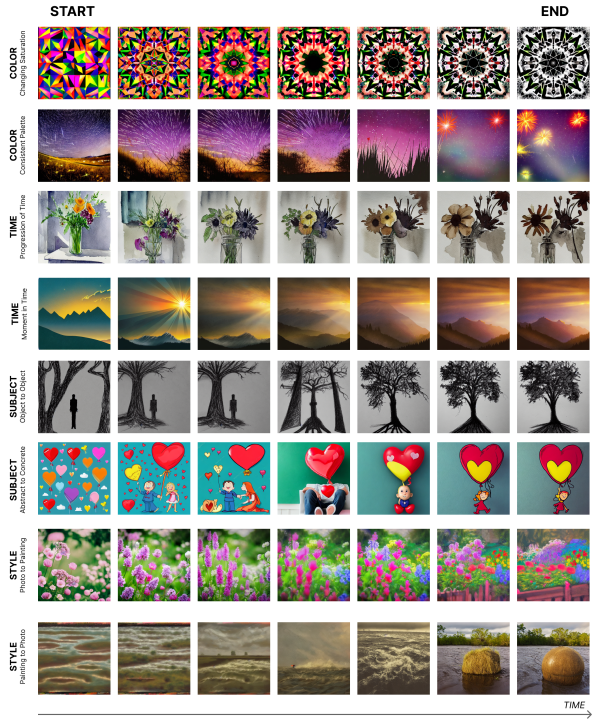

Figure 4. Types of transitions (color, time, subject, style) participants created to represent visual change.

a music artist, described how even though they had no professional experience in visual work, one way they understood music was as changes in color and visual intensity.

> *"Songs can kind of sound brighter or darker, depending on what sound you use or the frequency range that's dominant. You can think about how color might inform our understanding of this sound, like frequency range of darker colors mapping to shorter wavelengths on a color spectrum." - P8*

Using Generative Disco, P8 was able to act on this notion. They brought in an original composition and had an artistic vision for what they wanted their video to achieve. They wanted to express vibes of by sunlight streaming through a window, morning coffee, and yellows and greens to evoke hiking imagery. To achieve their vision, they utilized the visual brainstorming pipeline to find generations representing *"Golden sky, sun emerging slowly, sunny, vignette, warm, storybook"* and *"Cozy cabin basking in sunrise, vignette, warm, storybook illustration"*. A time transition generated with the golden sky image and sun rays (Fig. 4-Row 4) helped them express those morning sentiments and the sense of musical brightness they referred to.

P7 was a classically-trained musician who was also a novice to video creation. They found it easy to grasp generative workflow concepts like hyperparameters and use the visual brainstorming pipeline to explore conceptual spaces like color and symbols. They explored different instantiations of heart concepts (Fig. 4-Row 6) to pull in the color red and symbolize love for the lyric *"with somebody who loves me"*.

> *"I am really impressed by it. I think it's so cool, because I don't have an animation background or a computer science background. So having this interface–as someone who has no experience in either of those fields–it was very user-friendly, really fun to experiment with. How I was able to create a final product with an idea that I had in my brain–with no experience–I didn't think that that was possible."* -P7

Video professionals could also create in a music-first way. Rather than considering music as just one track or as an underscoring background element, they could engage with the different structures latent in the songs. They could move across the layers in the music, moving from beats to lyrics to other musical elements highlighted by amplitude changes in the waveform. For example, P2 first captured a quick succession of instrumental notes with short subject transitions and then captured longer musical phrases (a crescendo and a heavy metal breakdown) with a burst of color and style change. This is pictured in the second to last interval of Fig. 3, where a grayscale style transforms into a psychedelic one.

It is worth noting that some freelancers had blended experience with music and video (P1, P4, P10). For these participants, a primary strength of Generative Disco was that it computationally assisted with audiovisual alignment, which is something that is difficult to manually achieve. Freelancers drew upon their expertise in color and motifs to make the visuals resonate with the music at a high-level, while the system also handled the way visuals would snap to the music at the low-level. P10 appreciated this audiovisual alignment support as they visualized a locked groove from techno music. They used color as a strategy to highlight percussive elements in the music.

> *"[I am] trying to keep to the philosophy of what a locked groove is, making small variations between the broader theme, from black and white to color." -P10*

3.2.4. Failure Modes (RQ3) Motion artifacts were one of the main deciders for what made a generated interval usable or not. For example, P5 once tried to generate an interval depicting a couple. In the middle, intermediate frames pictured extra people, introducing a glitchy motion artifact. A similar issue was jitter; because our text-to-video inference process was conditioned on the input music, when the music had too many simultaneous features, the generated video could suffer from excessive audioreactivity.

Additionally, our system implementation was based on a model that did not afford composition control or camera motion control. Faces and bodies were prone to distortion when interpolated, causing participants to stay away from picturing people. Without camera motion control, participants were unable to take actions that are ingrained in their usual workflows such as panning, zooming, and rotation. We can incorporate new advancements from text-to-video models to implement features that resolve these issues.

4. Discussion

4.1. Digital "Double Hatting"

The world is more multimedia than ever. A musician who wants to release a song needs album art and video visualizers to help visually package their work on social media. The shortform content creator who wants to tell stories needs image and video assets to make their stories more immersive.

Many existing tools give people the means to create visual content if they are willing to learn tool-specific skill verticals (e.g. Photoshop, Premiere). When users learn the design languages specific to domains like music or video, they often find that the artistic expertise they developed in one domain does not carry over to the next. For example, a musician may work on mixing and mastering compositions within a music software, but not be able to translate that deep understanding of their music into design actions in a video editing software, because they have different low-level tool primitives. The language interactions in Generative Disco pulled people out of that low-level focus so that they could focus on the high-level story they wanted to craft around music. This form of interaction proved easy to learn: all participants learned how to utilize every generative function of Generative Disco within fifteen minutes, which is vastly different from the steeper learning curves expected to master audiovisual tools.

We show in our study that Generative Disco could enable *horizontal* expansion across different domains and increase skill mobility. This result adds a dimension to previous findings that generally only capture vertical acceleration domain experts get from generative AI in the form of speed and productivity gains (Noy & Zhang, 2023). It also adds a datapoint that runs counter to concerns raised about AI deskilling.

4.2. Multimodal Expression of Vibe

Generative Disco helped users express abstract concepts like musical vibe. Participants drew upon their artistic expertise to apply color and symbols that made their music visualization less literal, more narrative, and capable of accessing higher-order expressions of aesthetics and mood. We found that there were certain conceptual spaces such as color and emotion that participants kept returning to as ways of bridging the music and visual modalities. Discovering more multimodal spaces and ways of exploring them can help generative workflows better leverage artistic expertise. Creating richer points of interaction in these tools will give creative professionals more ways to highlight the creativity they are fundamentally hired for.

4.3. Future Work and Limitations

Generative Disco can still be improved on both its music and visual dimensions. We plan to explore methods for automatic music segmentation by analyzing songs in terms of melody, lyrics, dynamics, and rhythm. This can help users build longer narratives by reducing their listening taskload. Furthermore, we plan to incorporate more control modalities and stronger motion priors (Guo et al., 2023; Zhang & Agrawala, 2023).

Generative Disco's characterization of digital double hatting can continue to be studied under different sociotechnical lens. This form of horizontal skill expansion can impact transparency within client-artist relationships. Additionally, participants made observations that there could be impacts on labor ecosystems tertiary to their own, such as the stock footage space. Digital double hatting could also potentially introduce friction between creatives, who may have previously depended on each other through collaboration and joint projects. The longitudinal impacts of generative AI on the social dynamics surrounding creative professionals merits future work.

5. Conclusion

In this paper, we introduce Generative Disco, a generative text-to-video workflow for music visualization. In a user study field testing Generative Disco with audio and visual freelance creatives, we found that creatives found the generative tool easy to

use, expressive, and capable of supporting a number of professional use cases. Generative Disco helped creative professionals work across modalities and empowered them to have more artistic reach. We characterize type of labor augmentation as digital "double hatting" and detail how it can increase the skill mobility of creative professionals into adjacent domains.

References

Artlist. (2024). *Artlist business trend report*. https://artlist.io/blog/trend-report/

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., & Ramesh, A. (2024). Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators

Brüns, J. D., & Meißner, M. (2024). Do you create your content yourself? using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. *Journal of Retailing and Consumer Services*, *79*, 103790. https://doi.org/https://doi.org/10.1016/j.jretconser.2024.103790

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023, April). *Generative ai at work*. https://doi.org/10.3386/w31161

Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., Le Khac, P., Melas, L., & Ghosh, R. (2021, July). Dalle mini. https://doi.org/10.5281/zenodo.1234

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models.

Gero, K. I., Liu, V., & Chilton, L. B. (2021). Sparks: Inspiration for science writing using language models. https://doi.org/10.48550/ARXIV.2110.07640

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., & Dai, B. (2023). Animatediff: Animate your personalized text-to-image diffusion models without specific tuning.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022). Imagen video: High definition video generation with diffusion models.

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3501825

Liu, V., Qiao, H., & Chilton, L. (2022). Opal: Multimodal image generation for news illustration. https://doi.org/10.48550/ARXIV.2204.09007

Long, T., Gero, K. I., & Chilton, L. B. (2024). Not just novelty: A longitudinal study on utility and customization of an ai workflow. *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 782–803. https://doi.org/10.1145/3643834.3661587

Lyu, Y., Zhang, H., Niu, S., & Cai, J. (2024). A preliminary exploration of youtubers' use of generative-ai in content creation. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613905.3651057

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Stable diffusion videos. (2022). https://huggingface.co/stabilityai/stable-diffusion-2

Upwork. (2023). *Freelance forward 2023*. https://www.upwork.com/research/freelance-forward-2023-research-report

Wang, S., Menon, S., Long, T., Henderson, K., Li, D., Crowston, K., Hansen, M., Nickerson, J. V., & Chilton, L. B. (2024). Reelframer: Human-ai co-creation for news-to-video translation.

Wang, S., Petridis, S., Kwon, T., Ma, X., & Chilton, L. B. (2023). Popblends: Strategies for conceptual blending with large language models.

Yoganarasimhan, H. (2013). The value of reputation in an online freelance marketplace. *Marketing Science*, *32*(6), 860–891. https://doi.org/10.1287/mksc.2013.0809

Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models.