I N S Δ M

JOURNAL OF CONTEMPORARY MUSIC, ART AND TECHNOLOGY

Computing Short Films using Language-guided Diffusion and Voc-oding through Virtual Timelines of Summaries

Luís Arandas, Miguel Carvalhais and Mick Grierson

INSAM Journal of Contemporary Music, Art and Technology
No. 10, July 2023, pp. 71–89.
https://doi.org/10.51191/issn.2637-1898.2023.6.10.71

INSAM

DOI https://doi.org/10.51191/issn.2637-1898.2023.6.10.71

Luís Arandas*
University of Porto – INESC-TEC,
Porto, Portugal

Miguel Carvalhais**
University of Porto – i2ADS,
Porto, Portugal

Mick Grierson***
University of the Arts London – CCI,
London, United Kingdom

COMPUTING SHORT FILMS USING LANGUAGE-GUIDED DIFFUSION AND VOCODING THROUGH VIRTUAL TIMELINES OF SUMMARIES[1]

Abstract: Language-guided generative models are increasingly used in audiovisual production. Image diffusion allows for the development of video sequences and some of its coordination can be established by *text prompts*. This research automates a video production pipeline leveraging CLIP-guidance with longform text inputs and a separate text-to-speech system. We introduce a method for producing frame-accurate video and audio summaries using a virtual timeline and document a set of video outputs with diverging parameters. Our approach was applied in

* Author's contact information: luis.arandas@inesctec.pt.
** Author's contact information: mcarvalhais@fba.up.pt.
*** Author's contact information: m.grierson@arts.ac.uk.

1 The research leading to these results was conducted at the UAL Creative Computing Institute (03-08/2022) and financially supported by the Portuguese Foundation for Science and Technology (FCT), through the individual research grant 2020.07619.BD and by the project "Experimentation in music in Portuguese culture: History, contexts and practices in the 20th and 21st centuries" (POCI-01-0145- FEDER-031380), co-funded by the European Union through the Operational Program Competitiveness and Internationalisation, in its ERDF component, and by national funds, through the Portuguese FCT.

the production of the film *Irreplaceable Biography* and contributes to a future where multimodal generative architectures are set as underlying mechanisms to establish visual sequences in time. We contribute to a practice where language modelling is part of a shared and learned representation which can support professional video production, specifically used as a vehicle throughout the composition process as potential videography in physical space.

Keywords: artificial filmmaking, deep generative models, language-guided diffusion, short film computing, audiovisual composition, multimodal sequencing.

Introduction

Deep generative models have been used in film and audiovisual production by producing data according to a learned representation (Akten, Fiebrink, and Grierson 2020). Variational autoencoders and adversarial networks proved to be very efficient in generating both video and sound sequences, and more recently diffusion has become of great relevance to the state of the art (Dhariwal and Nichol 2021). Outputs are generated from deep generative models at least in two types of procedures: 1) to sample a compressed learned representation based on feature values captured by its learning procedure, e.g., a latent space vector; and 2) to reconstruct specific missing pieces of data according to their compressed learned representation, which is a discrete procedure and works as a filter that practitioners can control (Brooks, Holynski, and Efros 2022). Natural language has become a guiding principle in each one of these procedures, namely in image diffusion, where *text prompts* can sample a model and guide arbitrary data reconstruction, e.g., *inpainting* and *outpainting* (Chang et al. 2023). With implementations that work out moving images, the diffusion process can be established as continuous and sequences of prompts act at specific frames (Liu and Chilton 2022b).

We propose to consider text as the main representation mechanism of a set of generative models by building on image diffusion implementations with classifier guidance (Offert 2022). Ways of defining event sequences as sets of text prompts are tied to specific frames and this happens as models like CLIP evaluate every diffusion iteration and both text and image embeddings are scheduled first-hand (Kim, Kwon, and Ye 2022). We propose an implementation which wraps a virtual timeline organised with instructions for the entirety of each produced render. We design three possible sequencers in which the programmatic procedure of frame-by-frame production is automated by plugging an image diffusion system with a transformer summariser and a text-to-speech (TTS) vo-

coder where arbitrary texts can guide output sequences. In this way, a virtual timeline can control both the image and sound systems with a shared text representation, presenting a modular approach practitioners can build upon considering a model's experimental ability to create around specific initial inputs. A film using these technologies is documented and a working implementation is provided to compute sets of frames and audio buffers from text files using established deep learning libraries. By coordinating the used models together, we provide research on language-guided sequencing and establish a discussion on possible futures of filmmaking and audiovisual production using purely deep generative models which coordinate field of view representations with speech-based soundtracks.

1. Language-guided diffusion

Diffusion appears as a successor of other generative model architectures demonstrating remarkable results in vision (Croitoru et al. 2022). Inspired by non-equilibrium thermodynamics models, implementations work through a *forward* and *reverse* process, perturbing data using, e.g. Gaussian noise, and gradually learning to reverse it back (Sohl-Dickstein et al. 2015). By doing this procedure in steps, the approximation mechanism can be guided by natural language, each frame generating network encodings (Nichol et al. 2021). Diffusion models with classifier guidance allow us to generate images from *text prompts,* where a trained diffusion model score estimate is computed with a gradient of an independent image classifier (Dhariwal and Nichol 2021), as opposed to other latent sampling methodologies, see Rombach et al. (2022). When achieved using a network like CLIP (*Contrastive Language-Image Pre-training*), each diffusion step is guided towards a supposed natural language description, chunking each image to the model's compressed representation, often in 224 x 224 px (Radford et al. 2021). Without any conditional input mask, when developing a moving image entirely from a textual description, current implementations allow diffusion of a first frame from noise and further play with diffusion step percentages and variable prompts at future frames (Nichol and Dhariwal 2021).

1.1 Moving image model architecture

One possible way to successfully generate sequences of frames with classifier guidance with no image input relies on diffusing a first frame from a prompt and conditioning the next ones with arbitrary transformations and step skipping.[2] Flow coherency can then be worked out, allowing new images to visually

2 There are very different ways to point the first diffusion iteration towards some visual

match the previous while incrementing new model seeds. Simple image diffusion architectures can automate the process of creating *image-text* embeddings from arrays of prompt strings and guide the denoising progressive sampling (Song, Meng, and Ermon 2020). From the automation of this process, compensation can be added from the second frame onwards, passing the resulting image data as a texture to separate models such as, e.g. MiDaS (Ranftl et al. 2022) and AdaBins (Bhat, Alhashim, and Wonka 2021), estimating depth maps and computing projections on controllable fields of view using renderer cameras. Using this procedure each returned mask is used with specific weight in the next diffused frame, allowing to practice moving image flows; ways of introducing text prompts at specific frames have already been successful in coordinating CLIP embeddings with the process of progressive sampling (DDIM) (Salimans and Ho 2022). Recently introduced implementations with depth transformation have emerged with the notion of artificial camera objects creating a field of view in the processing workflow, resulting in rough spatial blueprints (Ravi et al. 2020);[3] as well as produced movement by estimating concrete separations between objects from each frame's emerging shapes.[4] With long diffusion renders with no variable prompt and low percentages of step skipping, saturation is experienced for the lack of variability in the picture. Current frames here are understood as enforcing future ones towards new variations and bridges with *video-to-video* rationale (Kim, Kwon, and Ye 2022).

2. Extending diffusion architectures with timelines

Implementations like DALL-E 2 and Imagen have had great impact in exposing how generative models can create realistic images from a description of natural language (Ramesh et al. 2022; Saharia et al. 2022), yet open implementations offer great modularity in integration e.g., Katherine Crowson's guided diffusion, disco diffusion[5] and stable diffusion; of classifier guidance (Dhariwal and Nichol 2021). Each video sequence starts by initially establishing sets of

structure with a single image condition, in the provided implementation we create a 705x384 black pixel array as an option, perceptible in section 3 using high step skipping.

3 Acute motion changes between diffused frames are perceptible by pixel clamping created at the direction of the camera transform. In our implementation we zoom and rotate both x and y axes using a fixed angle across the sequence.

4 The potential for future work in this area should be acknowledged, allowing for the introduction of other image recognition and segmentation systems to operate at the diffusion pass without the need to retrain a new system.

5 Our implementation targets almost every released version of the *disco-diffusion* open-source project, where we integrate the virtual timeline and speech system to be used on consumer-grade GPUs.

text prompts at specific frames and render towards a max number to then encode at variable frame rates; e.g., a simple automation of CLIP guidance is to incrementally establish sets of key frames with noise prompt values randomising embeddings. In our research we target the automation of this step, encoding into the CLIP model processed pieces of arbitrary text and derive them as values to a central timeline scheduler, with data structure and three proposed sequencing algorithms described in the high level processing diagram (Figure 1). Using an LED (*Longformer Encoder-Decoder*) transformer trained on longform texts with a separate GPT-J-6B,[6] we compute sets of summaries and schedule them using a simple matrix with a backbone of *frame:prompt* (Beltagy, Peters, and Cohan 2020), invoking both target embeddings and specific narration events in a separate audio track, with speech derived from both transformers. Working with sections is also possible, as the length of both modalities is of relevance when matching and with our working implementation variability can be further explored as well as, e.g., quantisation by incrementing values on the step skipping across diffused frames, as a way in which different shots and model-specific variables can be organised with their own language sections in higher sequence length.[7] To provide grounds for appropriation we set up three simple sequencers to define working timelines based on: 1) an arbitrary input text file; and 2) the computed speech length or user-defined length. The max number is a variable derived from the entire summary set over the image diffusion process regardless of derived prompts, computing at 25 frames-per-second (FPS) and dependent on sequencer choice, with 44.1kHz using spectral masking on each decoded audio batch.

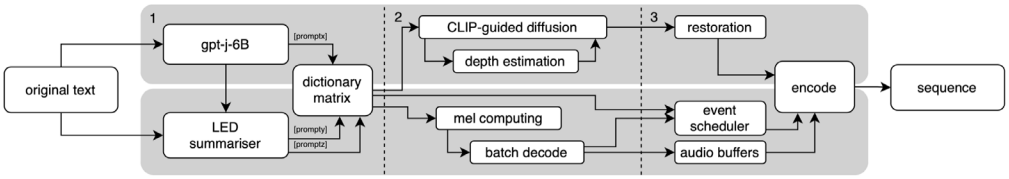
2.1 Sequencing audio buffers and video frames

Transforming the input text through a timeline to serve as a session template of the generative model architecture is here proposed as a methodology to sequence the already working integrations of language-guided diffusion, depth computing and TTS. We propose a programmatic procedure whereby feeding text to a timeline sequencer we set up a matrix storage, *frame:prompt* dictionaries, generated text and other needed variables regarding audio buffers and diffused frames. In Figure 1, the processing workflow is divided from the original text string input to the sequence output, where from left to right: 1) virtual timeline definition and text processing through an LED summariser trained on

6 Third person reference and paragraph-sized coherence in the produced speech is characteristic of summarisation, yet we implement a transformer generator from EleutherAI to add variability between image and TTS prompts.

7 Previous work has been proposed on trying to model sequences such as film dialogue using neural networks, see Sunspring (2016) by Oscar Sharp.

BookSum dataset (Kryscinski et al. 2021) and a GPT-J-6B instantiation (Muennighoff 2022), where the LED computes both the raw summary and a GPT-generated one using the raw as input; 2) image diffusion with depth transformation and TTS; and 3) the encoding process, where timeline variables can be further used for editing. The timeline object has a matrix where the sequencers establish the entire prompt sequence and their specific frames, depicted in Algorithm 1, can be understood as: 1) based on sequence length from user input, space prompt summary outputs with equal distances at defined FPS and compute their TTS separately to the same folder; 2) without sequence length definition calculate sequence length from TTS in seconds and compensate one each side; and 3) without sequence length definition do step 2) but compute TTS separately and establish five second prompts for each summarised sentence. Matching speech length and diffused frames opens up future research possibilities regarding variable skip steps and how to organise transitions, which can be word accurate.[8] With these simple sequencers the timeline object can interface with CLIP



Algorithm 1: Populate timeline

Input: Raw text file $\mathcal{T}$, length $\mathcal{L}$
Output: Matrix $\mathcal{M}$ with params and keyframes

for $t \in \mathcal{T}$ **do**
  $\mathcal{M} \leftarrow \begin{bmatrix} [LED(t)] & [GPTJ(t)] \\ [[i_0 - i_n] & [p_0 - p_n]] \end{bmatrix}$
  for $i \leftarrow 0$ **to** $|\mathcal{M}_{1,1}|$ **do**
    $\mathbf{seq_1} \leftarrow [0, \frac{\mathcal{L}}{|\mathcal{M}1,1|}, 2 \cdot \frac{\mathcal{L}}{|\mathcal{M}1,1|}, ..., (|\mathcal{M}1,1| - 1) \cdot \frac{\mathcal{L}}{|\mathcal{M}1,1|}]$;
    let $\mathcal{L} \leftarrow \sum t_i$, where $t_i = TTS(M_{1,1}[i])$ for $i = 0, 1, ..., |\mathcal{M}_{1,1}| - 1$
    $\mathbf{seq_2} \leftarrow [0, \frac{\mathcal{L}}{\sum t_i}, 2 \cdot \frac{\mathcal{L}}{\sum t_i}, ..., (\sum t_i - 1) \cdot \frac{\mathcal{L}}{\sum t_i}]$;
    $\mathcal{M}_{2,2} \leftarrow \mathcal{M}_{1,1}$;
    let $\mathcal{L} \leftarrow \sum t_i + 5 \cdot (|\mathcal{M}1,1| - 1)$, where $t_i = TTS(M1,1[i])$ for $i = 0, 1, ..., |\mathcal{M}_{1,1}| - 1$
    $\mathbf{seq_3} \leftarrow [0, \frac{\mathcal{L}}{\sum t_i + 5 \cdot (|\mathcal{M}1,1| - 1)}, 2 \cdot \frac{\mathcal{L}}{\sum t_i + 5 \cdot (|\mathcal{M}1,1| - 1)}, ..., (\sum t_i + 5 \cdot (|\mathcal{M}1,1| - 1) - 1) \cdot \frac{\mathcal{L}}{\sum t_i + 5 \cdot (|\mathcal{M}1,1| - 1)}]$;
    $\mathcal{M}_{2,1} \in \{seq_1, seq_2, seq_3\}$;
  end
end
return $\mathcal{M}, \mathcal{L}$;

Figure 1. High-level diagram of the proposed architecture extension, where *timeline matrix* represents needed sets of values which guide both the image diffusion setup and vocoder TTS. In the algorithmic table we propose a matrix with processed outputs from the input text and sets of arrays with frame-prompt pairs, with the sequence length defined or not by the user.

8 Systems design specifically on grammar understanding with language models, see e.g., Chung et al. 2022.

embeddings (Kim, Kwon, and Ye 2022). Inference through tacotron2 is done separately from single summary elements and decoding the mel results using hifi-gan we establish timings at the beginning of each render, focusing part 2) of the processing workflow solely on diffusion and audio buffer generation (Shen et al. 2018; Kong, Kim, and Bae 2020).

3. Irreplaceable Biography (2022)

Film and audiovisual production has benefitted from these different types of generative models both in coordination and in isolation, for offline and real-time scenarios (Navas 2017). Shifting towards a realm where language dictates their sampling and approximation procedure, each model in a set works together to form objects in fields of view exposing a learnt representation, with added fuzziness through implementation policies when computing a loss or score. For moving images, it is not just the textual guidance of the image frame that matters, but also how guidance keeps happening, and what to enforce when undesirable elements are represented in the same space that practitioners want to describe. Swapping models of the same target function in architectures such as the one proposed here is itself an act of conditioning, in what is supposedly a future of filmmaking and audiovisual production engineered around the human head. With that change, e.g., different voices can be used in the same time settings with the same text, representing cultural and physical nuances in the same way an image dataset does as the only visual record limiting a supposed memory system.[9] As a practical composition using the described methodologies, the film *Irreplaceable Biography* was produced around the premise of limit and established borders of representation a deep learning model can have. That if building sequencers in the current clock-based computers used to process image data, the possible outputs of complex vision systems are limited in the same way a human is, regarding the ability to translate or universalise beyond perspectivist context, harnessing a specific future which pursues the development of AI systems *as if* it was seen, imagined or thought, where learned representations keyframe specific cultural and human time. The architecture was designed to coordinate a 4-minute 20-second sequence of the poem *A Song of Myself* (1892 version) exposing what it could generate from summarisation procedures.[10]

9 Earlier research has demonstrated how generative vision models can bias a previously encoded image from their inner specific representation targeting *style-transfer* techniques (Akten, Fiebrink, and Grierson 2019).

10 Practical differences from the provided implementation rely on; 1) a secondary ImageNet model to yield better results, class-conditioning the diffusion backward steps for additional guidance; see Nichol et al. (2021); 2) forward tracking shot that doesn't saturate, stuck at 50% of

Figure 2. *Irreplaceable Biography*, 24 exported frames with semi-equal distance.

Results and future work

Text prompt sets have proven to be successful in combining single prompt weights through engineering specifications (Liu and Chilton 2022a). Diffusion renders can be automated and scheduled using templates which specify high-level parameters, namely frame-accurate sets of descriptions. Building on the already successful implementations of CLIP-guided diffusion with depth computing, we propose to automate the generative process with timelines where summaries prompt the full length of each sequence and provide a matrix representation to be used in different attempts. These summaries are dependent on user input, namely a text file, and determine visual elements of the video frame sequence and the TTS system at the same time; even if not literally matching. We provide an implementation using open machine learning standards where

condition from previous frame.

new sequences can be produced even with multiple runs at the same time. Using these methods, we rely on the success of both step-based diffusion coherence across frames, namely the roughness of each transform according to the first, with the depth estimation success. We propose a timeline template as a way to share and organise variables which condition the overall generative process. Natural language has been a fundamental way to declare all sorts of events in film and audiovisual practices, be it through poems, scripts, or just, in this case, patterns of written instructions which dictate future behaviour descriptively on a multimodal shared architecture of instanced network graphs. We promote that future work should derive new methods of coordination between trained models which together create small parts of human visual experience, considering the full video sequence to propagate culture by the fact models once learned from a small set of records created around the physical world. We claim that with our procedures text prompt sets can be used as blueprints to a time-based control mechanism, specifically with transformer and diffusion models (Esser, Rombach, and Ommer 2021), and help reveal the intrinsic fragments of datasets marked in each model's learned representation.

4.1 Working implementation

Targeting platforms which democratise torch model instantiation we provide a small Python library with a virtual timeline object to structure the depicted model architecture using OpenCLIP with a single ViT-B-32 with batches of 4 cuts,[11] and a separate grammar corrector FLAN-t5 (Subramanian 2018; Schuhmann et al. 2022; Chung et al. 2022). We built for NVIDIA graphics cards, tested with PNY 3090 24Gb XR8 and Quadro P6000. The timeline is organised around: 1) the full length of each sequence, manually added by the user or computed from the generated TTS length; 2) the sequencers which output usable instructions for CLIP-guidance; 3) text processors to derive new text and populate storage; and 4) other render-related variables such as namespaces, formal ratios, frame buffers. Our implementation is modular and as objective future work we propose: 1) audio diffusion TTS with classifier guidance has already been proposed and could be further implemented (Kim, Kim, and Yoon 2022); 2) a more manageable way to deal with frame-shot composition, by indexing specific parts of each step with direct object placement alongside camera transforms;[12] 3) speech-related sound manipulation and voice translation directly in the used library; 4) tokens (words) and concepts to remove from each prompt set and neg-

11 In every iteration each cut of the image is compared against the prompt, offering a choice to accumulate gradients in batches.

12 Recent research has been successful in still images using latent models, see Ma et al. (2023)

ative weight management; and 5) manageable camera movement descriptions with visual declaration. We provide command line installation with a dedicated environment tested on Ubuntu 22.04 with instructions[13] and a set of videos with different lengths using a set of texts from online poetry platforms. If under 250 characters the whole text is given to the summariser module, otherwise just the beginning is. The videos experiment with different camera trucks and keyframes to help follow the embeddings and how imagery develops, in purely black and white contexts with slow movement velocities.[14] Variability on a frame-by-frame condition regarding camera and cinematic languages should be exercised and current AI architectures should encompass higher-level declarations over spe-

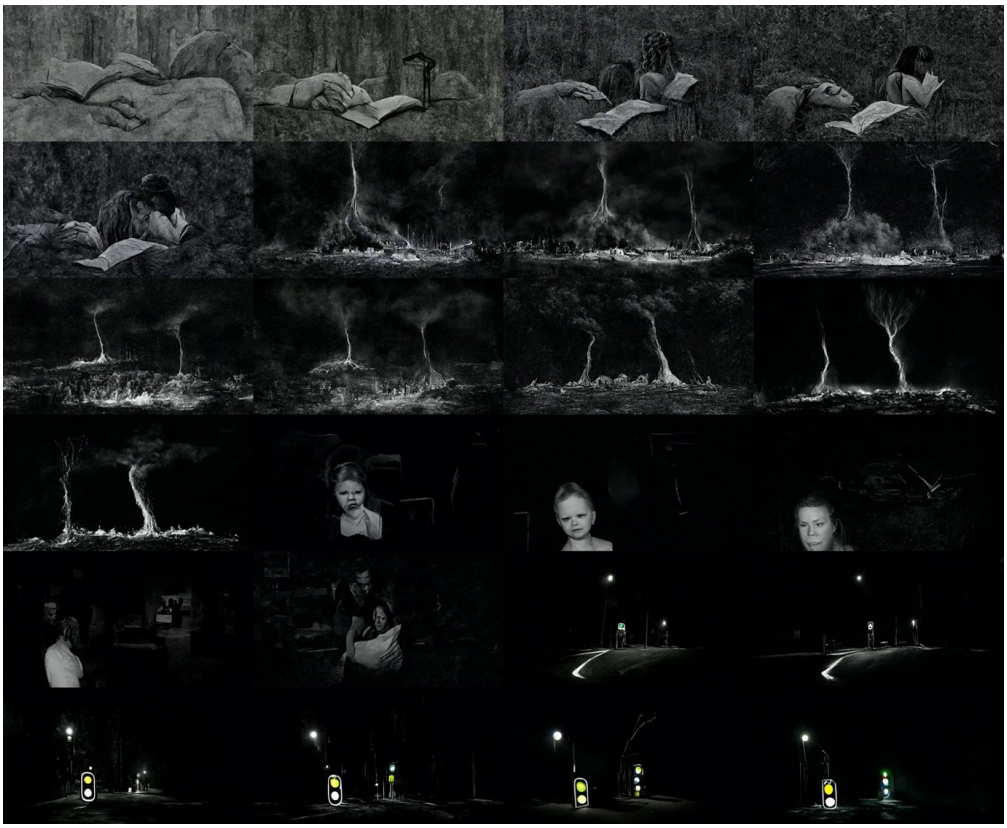


Figure 3. (30 seconds, 5 prompts) 4x6 frame export from the computed sequence using *Acquainted with the Night* by Robert Frost (1928), exercising diagonal translation. The generated prompts and timecodes are: 1) 00:00:00, "The narrator struggles to find clarity in the relationship between her and the other characters."; 2) 00:00:06, "Great scientists discover a natural disaster."; 3) 00:00:12, "Wife fear loss of job children."; 3) 00:00:18, "Eventually a stranger brought relief to discover time actually helps to bring back wife."; 4) 00:00:24, "The narrator continues to listen to the distortion of traffic lights and becomes annoyed when people fall asleep on the roads".

13 https://github.com/luisArandas/virtual-timeline-clipguided (accessed 10-07-2023).

14 The provided examples are not entirely greyscale and had no color correction whatsoever.

cific trajectories following realism standards. At each new prompt keyframe, the embedding computation can be injected with needed distributions.

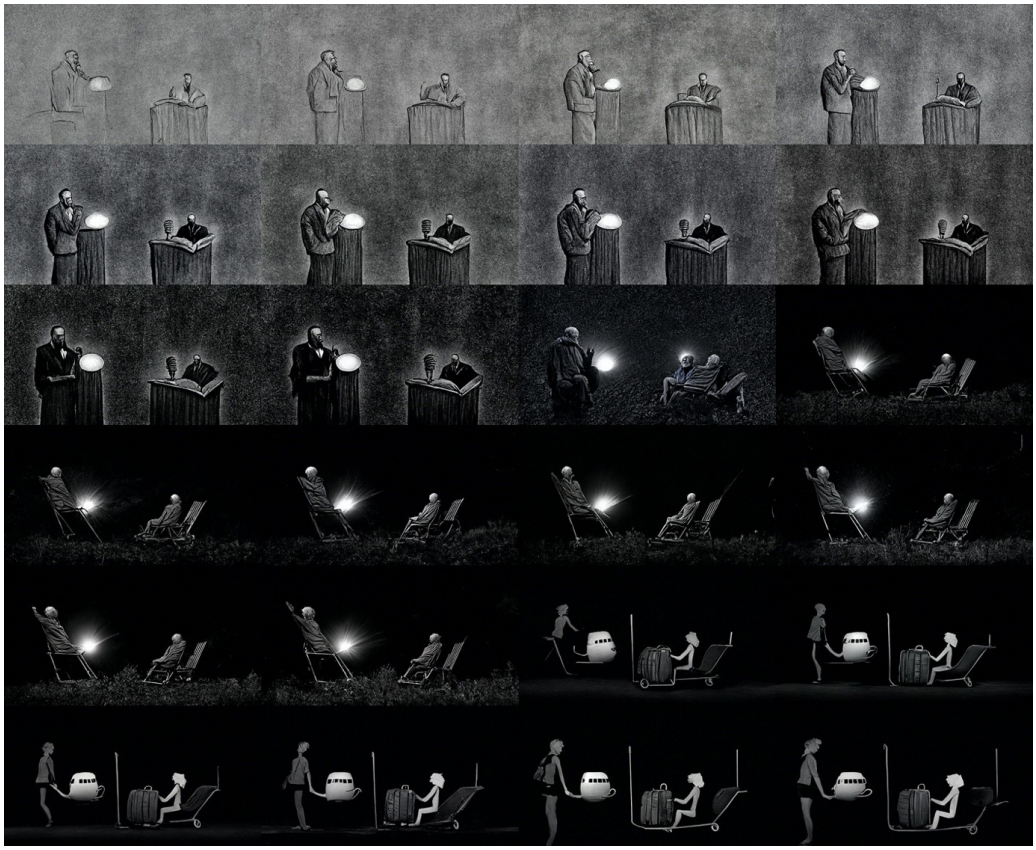


Figure 4. (30 seconds, 3 prompts) 4x6 frame export from the computed sequence using *Do not go gentle into that good night* by Dylan Thomas (1947), exercising slow movement. The generated prompts and timecodes are: 1) 00:00:00, "The narrator delivers a long speech in which he implores us to rage against the dying of the light and against mankind's tendency to go greedily into that good night."; 2) 00:00:10, "He lists all sorts of examples of people who have succumbed to the darkness and urged us to do the same: Old men who have spent their lives helping to keep the sun out of the sky"; 3) 00:00:20, "But who lost it because they were too late to help it travel its way".

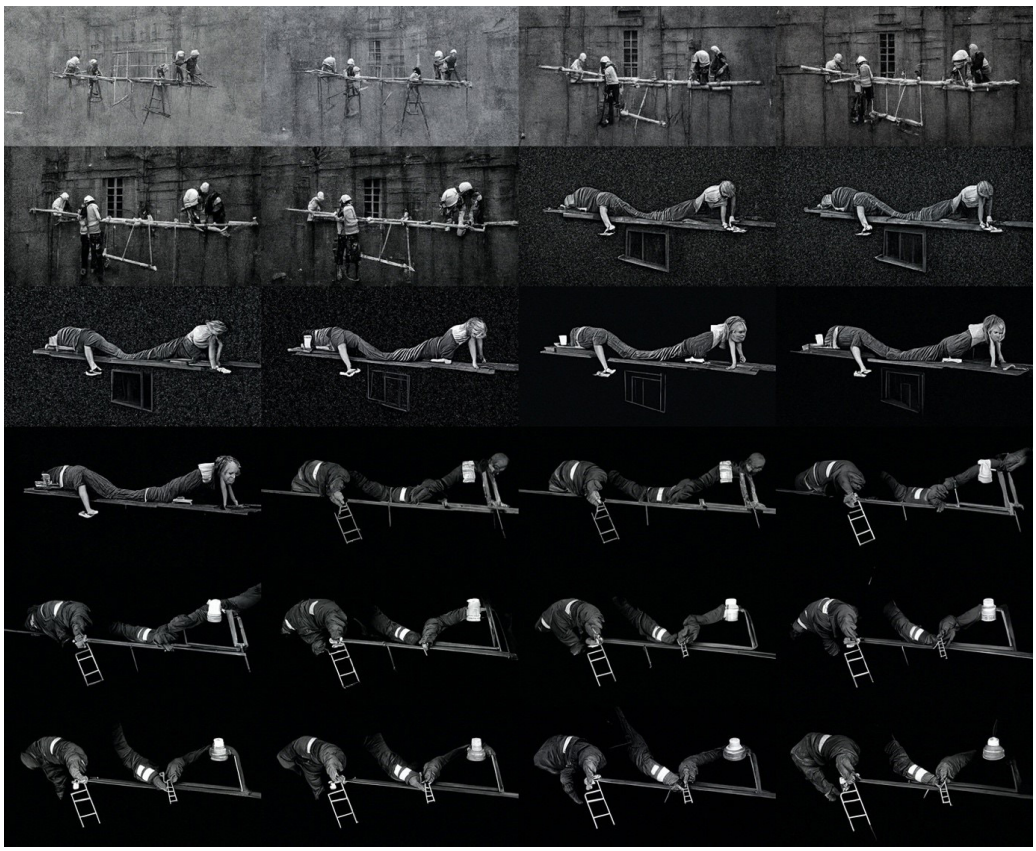


Figure 5. (30 seconds, 3 prompts) 4x6 frame export from the computed sequence using Scaffolding by Seamus Heaney (1966), exercising slow movement. The generated prompts and timecodes are: 1) 00:00:00, "The wall builders that we have here at Boulogne house are careful to test the scaffolding before they start building."; 2) 00:00:10, "Make sure planks won't slip at busy times."; 3) 00:00:20, "Secure all ladders and tighten bolted joints".

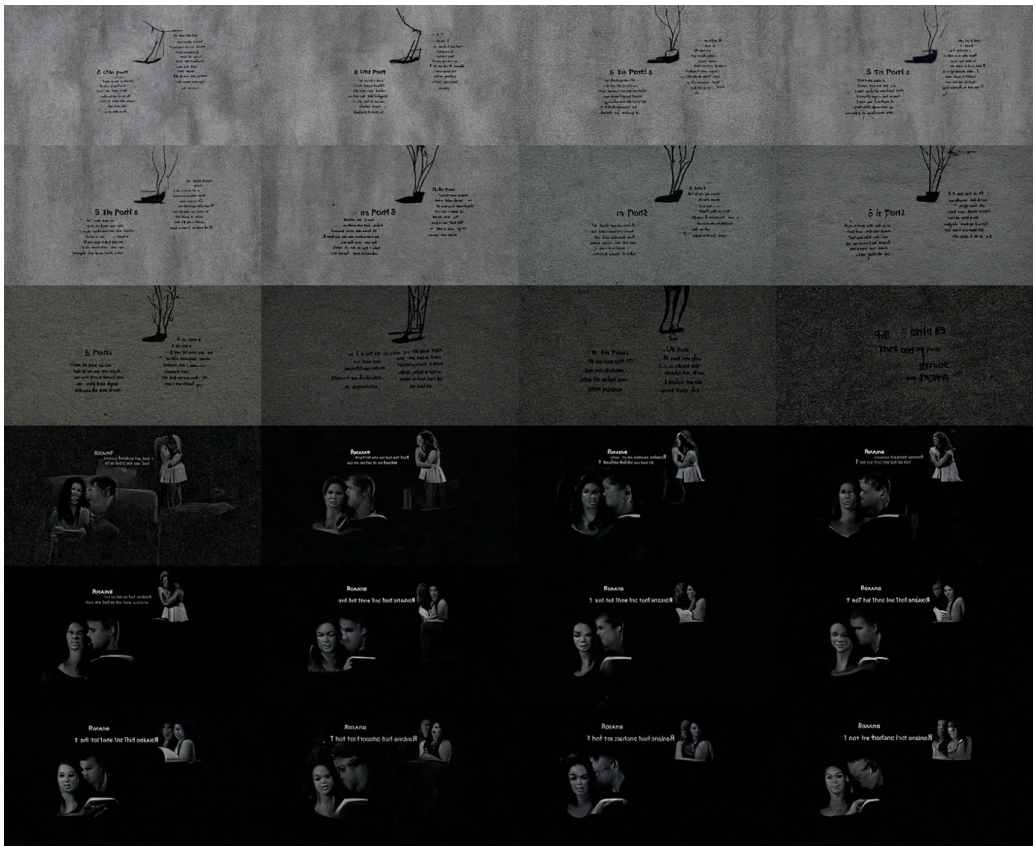


Figure 6. (20 seconds, 2 prompts) 4x6 frame export from the computed sequence using *Yours* by Daniel Hoffman, exercising visual dialogue design. The generated prompts and timecodes are: 1) 00:00:00, "In this short poem."; 2) 00:00:10, "Roxane reminds Christian that he is her true love and that without her he would be nothing".

Figure 7. (20 seconds, 2 prompts) 4x6 frame export from the computed sequence using *Still I Rise* by Maya Angelou (1978), exercising visual dialogue design. The generated prompts and timecodes are: 1) 00:00:00, "Still I will rise"; 2) 00:00:10, "Like a black ocean rising".

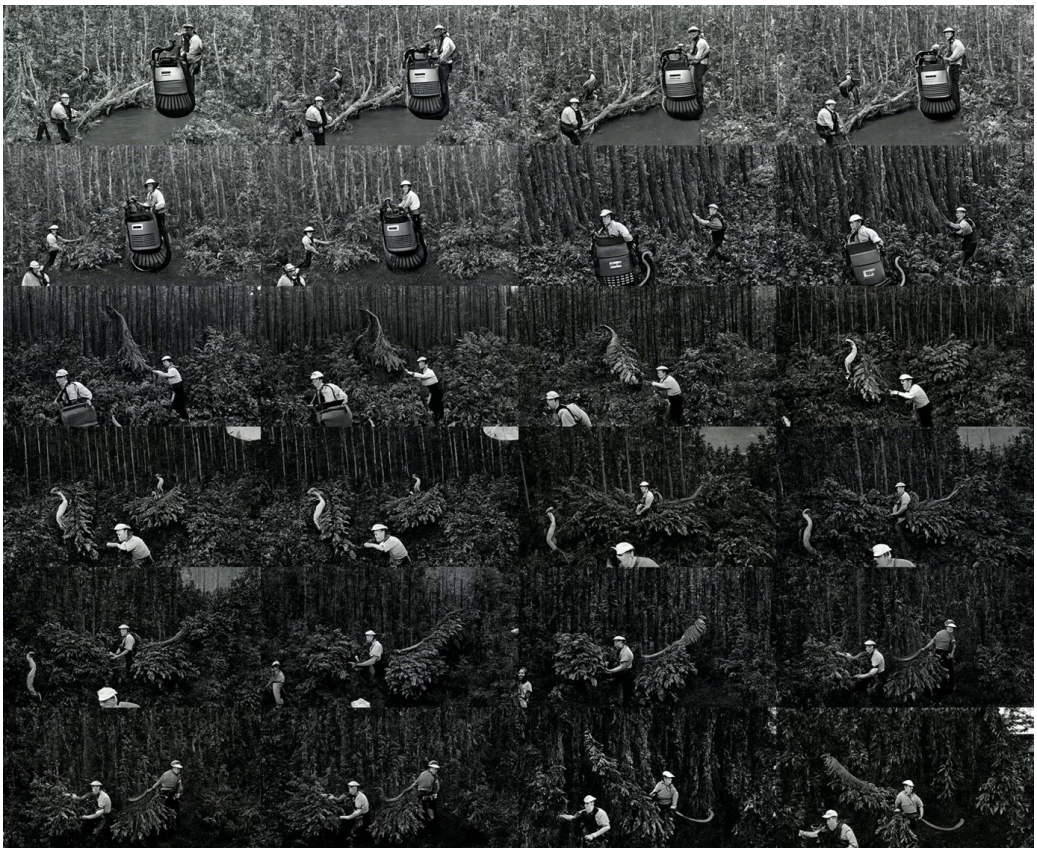


Figure 8. (25 seconds, 1 prompt) 4x6 frame export from the computed sequence using a piece of text extracted from a public website *The Royal Parks* about trees in Regent's Park and Primrose Hill. Selected paragraph: "Trees host complex microhabitats. When young, they offer habitation and food to amazing communities of birds, insects, lichen and fungi. When ancient, their trunks also provide the hollow cover needed by species such as bats, woodboring beetles, tawny owls and woodpeckers. One mature oak can be home to as many as 500 different species. Richmond Park is full of such trees, which is one of the reasons it has been designated a National Nature Reserve and Site of Special Scientific Interest". Selected prompt, used throughout the whole sequence: 00:00:00, "Narrator explains tree like oak import ecologist provide wide range biological service like Redmond park include wood butt wood owl wood pick beetle". This output has the transform on the *y* axis instead of *x* allowing positive rotation, exploring movement stability without enforcing new language without visual feedback.

Conclusion

Language has become a very successful method to coordinate image diffusion systems. From progressive sampling to still image editing mechanisms, diffusion models provide clear ways in which we can set up frame sequences. By extending language-guided video production we focused on how text prompts create templates of a programmatic procedure by defining event sequences and audio narration, proposing a virtual timeline implementation which allows us to compute arbitrary texts using established machine learning frameworks. By implementing a set of sequencers we designed a usable model architecture as an extension of image diffusion with classifier guidance where sets of prompts define both aesthetic and formal properties of the image and speech through generated summaries with frame accuracy. We experimented with the architecture's ability to represent a certain input text and how this coordination defines fields of view; which themselves are products of a model's approximation ability in representing the dataset it learned from. Our implementation follows open standards and allows us to produce new sequences using arbitrary text files with variable tracking shots. Further, we documented the short film *Irreplaceable Biography* as a material application of these methodologies and promoted a discussion on how learnt representations consequentially reveal aspects of physical reality by providing a rough simulation of what established film and audiovisual practices search for, outlining both documental and abstractive character. Systematic production of audiovisual sequences regarding text descriptions can benefit from simple implementations as diagrams of automatic procedures.

List of References

Akten, Memo, Rebecca Fiebrink, and Mick Grierson. 2019. "Learning to see: you are what you see". *ACM SIGGRAPH Art Gallery*: 1–6.

Akten, Memo, Rebecca Fiebrink, and Mick Grierson. 2020. "Deep Meditations: Controlled navigation of latent space". arXiv:2003.00910.

Beltagy, Iz, Matthew E Peters, and Arman Cohan. 2020. "Longformer: The long-document transformer". arXiv preprint arXiv:2004.05150.

Bhat, Shariq Farooq, Ibraheem Alhashim, and Peter Wonka. 2021. "Adabins: Depth estimation using adaptive bins". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Brooks, Tim, Aleksander Holynski, and Alexei A Efros. 2022. "Instructpix2pix: Learning to follow image editing instructions". arXiv preprint arXiv:2211.09800.

Chang, Huiwen, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, and Michael Rubinstein. 2023. "Muse: Text-To-Image Generation via Masked Generative Transformers". arXiv preprint arXiv:2301.00704.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, and Siddhartha Brahma. 2022. "Scaling instruction-finetuned language models". arXiv preprint arXiv:2210.11416.

Croitoru, Florinel-Alin, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2022. "Diffusion models in vision: A survey". arXiv preprint arXiv:2209.04747.

Dhariwal, Prafulla, and Alex Nichol. 2021. "Diffusion Models Beat GANs on Image Synthesis". In 34th Conference on Neural Information Processing Systems (NeurIPS 2021).

Esser, Patrick, Robin Rombach, and Bjorn Ommer. 2021. "Taming transformers for high-resolution image synthesis". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Kim, Gwanghyun, Taesung Kwon, and Jong Chul Ye. 2022. "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Kim, Heeseung, Sungwon Kim, and Sungroh Yoon. 2022. "Guided-tts: A diffusion model for text-to-speech via classifier guidance". International Conference on Machine Learning.

Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. 2020. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis".

Kryscinski, Wojciech, Nazneen Fatema Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R Radev. 2021. "Booksum: A collection of datasets for long-form narrative summarization". arXiv:2105.08209.

Liu, Vivian, and Lydia B Chilton. 2022. "Design Guidelines for Prompt Engineering Text-to-Image Generative Models". CHI Conference on Human Factors in Computing Systems.

Ma, Wan-Duo Kurt, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. 2023. "Directed Diffusion: Direct Control of Object Placement through Attention Guidance". arXiv preprint arXiv:2302.13153.

Muennighoff, Niklas. 2022. "Sgpt: Gpt sentence embeddings for semantic search". arXiv preprint arXiv:2202.08904.

Navas, Eduardo. 2017. "Machine Learning and Remix: Self-training Selectivity in Digital Art Practice". *Remix Theory*.

Nichol, Alex, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models". arXiv pre-

print arXiv:2112.10741.

Nichol, Alexander Quinn, and Prafulla Dhariwal. 2021. "Improved denoising diffusion probabilistic models". International Conference on Machine Learning.

Offert, Fabian. 2022. "Ten Years of Image Synthesis". lwerkstatt.org/blog/ten-years-of-image-synthesis (visited July 14, 2023).

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. "Learning Transferable Visual Models From Natural Language Supervision". arXiv:2103.00020.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. "Hierarchical text-conditional image generation with clip latents". arXiv preprint arXiv:2204.06125.

Ranftl, R., K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. 2022. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer". *IEEE Trans Pattern Anal Mach Intell* 44 (3): 1623–1637. https://doi.org/10.1109/TPAMI.2020.3019967. https://www.ncbi.nlm.nih.gov/pubmed/32853149.

Ravi, Nikhila, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. "Accelerating 3d deep learning with pytorch3d". arXiv preprint arXiv:2007.08501.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. "High-resolution image synthesis with latent diffusion models". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, and Rapha Gontijo Lopes. 2022. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding". arXiv preprint arXiv:2205.11487.

Salimans, Tim, and Jonathan Ho. 2022. "Progressive distillation for fast sampling of diffusion models". arXiv preprint arXiv:2202.00512.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, and Mitchell Wortsman. 2022. "Laion-5b: An open large-scale dataset for training next generation image-text models". arXiv preprint arXiv:2210.08402.

Shen, Jonathan, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, and Rj Skerrv-Ryan. 2018. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions". 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP).

Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. "Deep unsupervised learning using nonequilibrium thermodynamics". International Conference on Machine Learning.

Song, Jiaming, Chenlin Meng, and Stefano Ermon. 2020. "Denoising diffusion implicit models". arXiv preprint arXiv:2010.02502.

Subramanian, Vishnu. 2018. *Deep Learning with PyTorch: A practical approach to building neural network models using PyTorch*. Birmingham: Packt Publishing Ltd.

COMPUTING SHORT FILMS USING LANGUAGE-GUIDED DIFFUSION AND VOCODING THROUGH VIRTUAL TIMELINES OF SUMMARIES (summary)

Language-guided generative models are being proposed to deal with numerous tasks across audiovisual production. Of great relevance has been the application of image diffusion to generate frames from *text prompts* and develop video sequences, formally describing content and other visual properties already found in still image composition. Current implementations with classifier guidance take advantage of models which represent images and text descriptions in a shared space; with this research, we automate a video production pipeline using CLIP-guided diffusion, allowing the introduction of arbitrary longform text inputs, pairing with a text-to-speech (TTS) system. Using a virtual timeline implementation we produce sets of frame-accurate summaries providing a method which allows us to produce sets of video frames and audio buffers for reproduction. Compliant with system design, we document the production of the film *Irreplaceable Biography* where our proposal targets deep generative model architectures' ability in coordinating visual elements through CLIP-guidance, among other moving image transformations like depth estimation. We introduce sets of summaries as text prompts to display a structure and automatically define sections of both video sequences and soundtracks using the same input. With this research we extend on the role of language to guide film and audiovisual production with its declarative and descriptive role, organising events and other formal properties of both the picture and sound targeting purely artificial outputs. We contribute to a future of filmmaking and virtual production where multimodal generative architectures are used as sequencers, following a simulation of the human head, perspective and experience. It is within current ventures of language modelling that we argue it can contribute to a shared representation which dictates how a video should *look like* and help leverage a simulation of what could be videography in the physical space.

Article received: April 4, 2023
Article accepted: June 9, 2023
Original scientific paper