xCoAx 2023 11th Conference on Computation, Communication, Aesthetics & X
2023.xCoAx.org
Weimar, Germany

Irreplaceable Biography: Possible Futures of AI Filmmaking Regarding a Model's Filter of Reality

Luís Arandas
luis.arandas@inesctec.pt
University of Porto, INESC-TEC, Porto, Portugal

DOI 10.34626/xcoax.2023.11th.389

Natural language has become a guiding principle of deep generative models. Across film and audiovisual production *text prompts* provide continuous control of video sequences establishing several approximation mechanisms. This article exposes the methods taken in the production of *Irreplaceable Biography* (2022), a film that extends current approaches to image diffusion systems by automating the prompt process, which dictates a blueprint for how each video sequence should *look like*. An algorithm was developed that leverages an independent transformer summariser, enabling a generative approach to sequence management and description, but most of all, the introduction of arbitrary manuscripts to then derive frame-accurate instructions. Targeting outputs entirely conceived by generative models, we expand on their ability to capture aspects of physical reality, conditioned by how they resemble specific datasets used during training. In our procedure, both their success and failure are posed as a filter of compositional value, whereby their ability to approximate what is previously represented expose culture at a specific time, as each dataset records fragments of the human.

Keywords: Language-guided Diffusion, Deep Generative Models, Artificial Filmmaking, Audiovisual Production, Short-film Computing.

Introduction

Deep generative models have been used to create film and media by generating data according to a learned representation (Saharia et al. 2022). Previous research was successful in computing video sequences by, e.g., establishing trajectories in latent spaces, examining encoded feature values inscribed in the system (Akten, Fiebrink, and Grierson 2020). Aesthetic specificity is understood as images are configured around high-level representations, as classes or groups, and by sampling a model we compute a simulation of how something *could be*, constraining images with its architecture and used dataset. Language has become a fundamental lever of such methods and audiovisual production shifts towards *text prompts* as a guiding principle (Nichol et al. 2021). Longevity becomes of consideration, as when more than one trained model interacts together the limits of representation become shared (Radford et al. 2019), and with the process behind *Irreplaceable Biography* (2022), we describe how current image diffusion architectures, which are multimodal by nature when guided by language, can be extended to a realm of automation on what has been previously posed as declarative and many times conversational.[1] Introducing an independent summariser network at the beginning of each sequence renders separate longform texts that can be used to further practical coordination of deep generative model architectures; understood as reality representation mechanisms which generate media with an emerging bias, aligned with their specific learnt representation and by resembling human vision laws both in their architecture and in the visual outputs (Whittington, Warren, and Behrens 2021; Ye, Xue, and Lin 2021).

Language-Guided Diffusion

Diffusion has had repercussions in both still and moving images (Kim, Kwon, and Ye 2022). Through a forward-reverse process, an image is perturbed using noise (e.g. Gaussian) in steps and neural networks gradually learn to reverse that process (Dhariwal and Nichol 2021). Without image input, the networks approximate frames from noise towards a desired text string to the best of their ability (Yang et al. 2022) and as generative models, implement many procedures previously achieved with e.g., adversarial networks (Li et al. 2020). Yet tools available to artists and the general public are built around the prompt input just as in conversational AI systems, requesting to declare images with language tricks.[2] Different algorithms contribute to the whole video sequence and when developing moving images,

[1] Considered here a presupposed interface given current implementations of image diffusion with classifier guidance, the infrastructure built around it, and how we experience causality as practitioners and filmmakers when dealing with *text prompts*.

[2] The network CLIP was used to build several diffusion architectures guided by language, and integrates with our summariser proposal. Methodologies to process *text prompts* have appeared and previously recognised as *prompt engineering* (Radford et al. 2021).

temporal coherence is exercised as new methods tinker with: skipping diffusion steps; blending and warp (Ilg et al. 2017); and, in this case, depth computing to provide a virtual projection with spatial information of the generated imagery (Bhat, Alhashim, and Wonka 2021). These methods allow frame shot composition and audiovisual development, if working with sound.[3]

Language-guided diffusion can also be used to enforce text prompts on video frames and produce new frames, which allows generative models to be used as signal processors (Mital, Grierson, and Smith 2013). Generative models are successful in the art world and in audiovisual production working by this rationale, making way for what is known as *video-to-video* (Loftsdottir and Guzdial 2022), a procedure used when diffusing purely from language or black frames (*previous/next*). Artworks from contemporary galleries to motion picture festivals have been using the lens of a neural network on videos delineating AI's ability to reconstruct determined frame data towards its inner representation of what it is (Steyerl 2019), and by focusing on production with different manuscripts without visual condition (frame-input) diffusing the initial one from noise and working out new diffusion step percentages from that defines types of flow coherency (Saharia et al. 2022). We control the amount of diffusion steps added between each specific prompt alongside monocular depth estimation to simulate movement in angles of a discrete field of view (Ranftl et al. 2022), predicting the next frames towards the next prompt with morphological coherence with the previous ones; defined as string objects, prompts with frame pairs are read in each diffusion render.

Computing Short Films from Texts

Text prompts define video sequences as the main representation mechanism of a generative model architecture (Liu and Chilton 2022), and when trained on different datasets provide aesthetically divergent outputs even if practically used with the same procedures; a characteristic of learning compressed representations of specific data (LeCun, Bengio, and Hinton 2015). By being tied to a dataset, models have not only been polarised into contextual success or failure working by percentages and loss values, but also pushed further audiovisual production by their abilities to abstract and synthesise images and narratives, emerging from their training influence (Chourdakis and Reiss 2017). *Irreplaceable Biography* (2022) is composed using a CLIP-guided image diffusion system and automates an

[3] Still frame composition has been developing extensively with textual instructions, which supports moving image production in itself. Vectors of prompts define timelines, coordinating embeddings and properties of the CLIP guidance on each diffusion step; in our case multiple CLIP models are used at the same time. Extending this method to an audio TTS system, a secondary set of prompts can then be used on the same render (Brooks, Holynski, and Efros 2022; Popov et al. 2020).

independent transformer summariser on previously produced text (Beltagy, Peters, and Cohan 2020), trained on the *BookSum* dataset (Kryściński et al. 2021). The added model computes vectors of summaries working as prompts at specific frames and audio buffers of a speech vocoder, exhibiting a new layer of influence on top of the original manuscript (Shen et al. 2018). On top of that, we propose a layering algorithm to glue the produced media in a fixed frame length and encode it into reproducible formats, having the voice lead how long each section takes. Walt Whitman's *A Song of Myself* (1892 version) is used as input and schedules the image together with a narration resembling a monologue, while the whole sequence is a forward tracking shot where each element of the produced frames gradually develops and disappears.

Current image diffusion systems coordinate several models to display multimodal capabilities and configure an observable field of view from single diffused frames. Natural language presents itself as a meaningful mediator in that coordination, as it has been in both the film and audiovisual industries over the years (Clark 2022). Regardless of modality specifics, each model contributes to representing a sequence constrained by their ability to resemble datasets, being defined on how the initially provided text will be filtered. Each established coordination may end up in different types of films with a totally different flow of narration following our proposal, where the produced vectors of summaries which will guide audio and image frames are automatically mapped to specific timings for the whole duration. Virtually all outputs can be divergent aesthetically and capture parts of the embedded texts, having a deterministic procedure or not (Ramesh et al. 2022).[4] With this method, we contribute to recognising generative models to capture and consequently represent aspects of physical reality, superficially marked in their inner representation through the specific datasets used to train them. From this point, short films entirely conceived by generative models can be understood as able to represent aspects of the world, with an added layer of interest, as they are developed according to both a structuralist understanding of the human and its subjective visual and auditory experience (Mitchell 2006).

[4] By their architecture models can break down reproducibility, e.g., with causality on used seeds, and that is of high importance in production which differs from other latent implementations (Rombach et al. 2022).

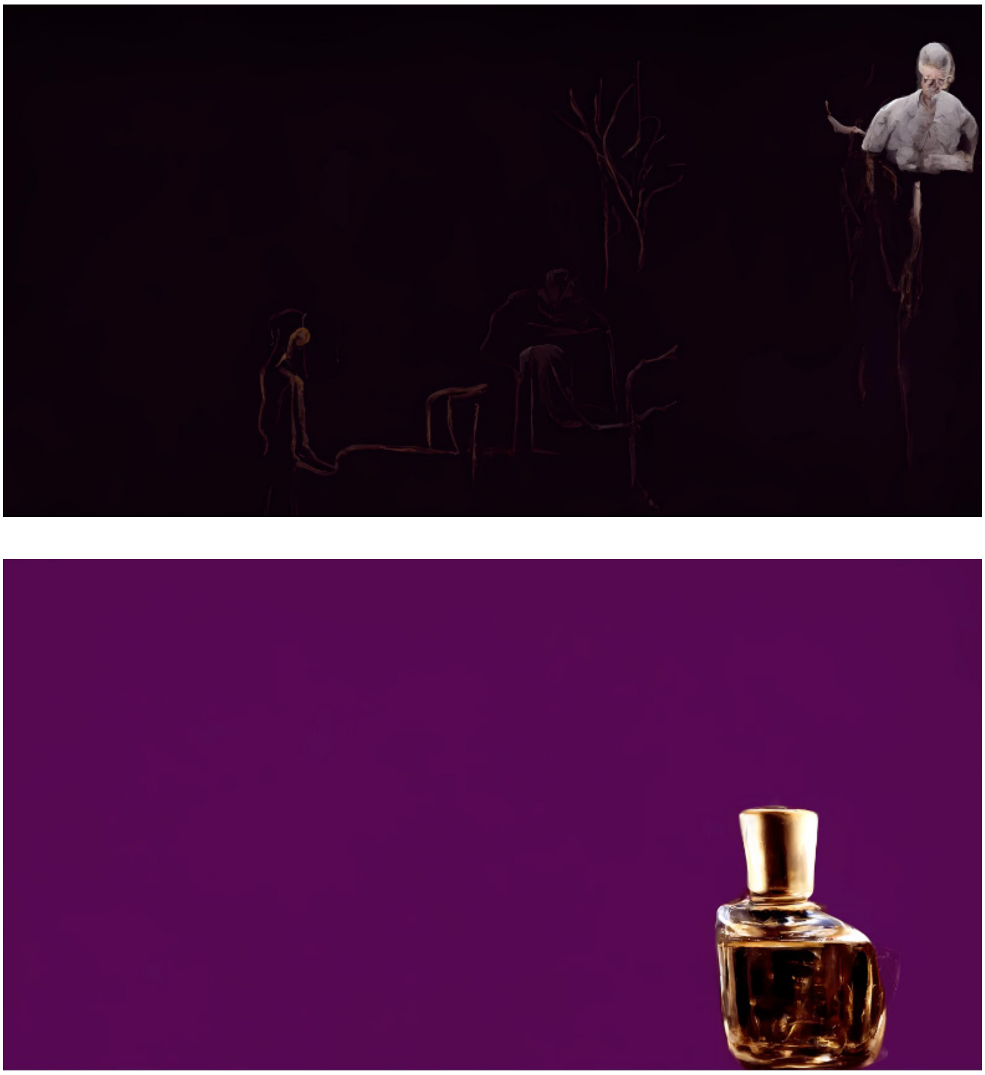

Figure 1, 2: Still frames from *Irreplaceable Biography* (2022) 04:35, and from Walt Whitman's *A Song of Myself* sequence (1892 version). 01:08 speech track: "The narrator explains that, at the age of 37, he is still healthy and strong and he hopes to live until death". 02:48 speech track: "The distillation of alcohol would intoxicate the narrator, but he will not let it intoxicate him for he believes that death would be preferable to this life of pure pleasure".

Conclusion

Language-guided generative models are being used across film and audiovisual production. Text prompts define outputs by approximating images towards their natural language description. Examining deep generative models as signal processors which capture aspects of the physical world at a specific time in their compressed representations of datasets, this article describes composition processes used in the film *Irreplaceable Biography* (2022) and methodologies which can be applied to compute new short films from arbitrary texts. By working on a field of view resembling an observable world, we propose a method to structure video sequences through vectors of summaries, extending a CLIP-guided diffusion system (Ravi et al. 2020). Visual artefacts develop and deform towards textual description alongside audio narration through synthesised speech, and the approximation mechanisms characteristic of deep generative models are here considered as providing a filter, defined by their ability to reconstruct each dataset, exposing society and culture by its bias, and resembling human visual experience. We theorise on how practical futures of filmmaking can benefit from such simulations and expose a methodology that practitioners can appropriate and build upon to compute new texts (Yang et al. 2022). Natural language guides a huge chunk of audiovisual art and specifically film practices, the role deep generative models play in resembling reality can be controlled by stepping back from a declaration prompts ask for, and finding ways to automate and direct their prediction (Navas 2022).

Acknowledgements. The research leading to this film 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. Supervision is provided by Prof. Miguel Carvalhais and Prof. Mick Grierson.

References

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. 2020. "Adabins: Depth estimation using adaptive bins." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4009-18. doi: 10.1109/CVPRW56347.2022.

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

Chourdakis, Emmanouil, and Joshua Reiss. 2017. "Constructing narrative using a generative model and continuous action policies." In *Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. Santiago de Compostela, Spain. doi:10.18653/v1/W17-3901.

Clark, Lynda. 2022. ''Towards 'Creativity Amplification': or, AI for Writers, or Beating the System'', *Writing in Practice,* no.7 (2022). issn: 2058-5535.

Dhariwal, Prafulla, and Alexander Nichol. 2021. "Diffusion models beat gans on image synthesis", *Advances in Neural Information Processing Systems,* 34: 8780-94. isbn: 9781713845393.

Ilg, Eddy, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. "Flownet 2.0: Evolution of optical flow estimation with deep networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1647-1655. doi: 10.1109/CVPR.2017.179.

Kim, Gwanghyun, Taesung Kwon, and Jong Chul Ye. 2022. "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426-35. doi: 10.48550/arXiv.2110.02711.

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

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep learning', *Nature*, 521: 436-44. https://doi.org/10.1038/nature14539.

Li, Bowen, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. 2020. "Lightweight generative adversarial networks for text-guided image manipulation", *Advances in Neural Information Processing Systems*, 33: 22020-31. isbn: 9781713829546.

Liu, Vivian, and Lydia B. Chilton. 2022. "Design Guidelines for Prompt Engineering Text-to-Image Generative Models." In *CHI Conference on Human Factors in Computing Systems*, A.No. 384: 1–23. https://doi.org/10.1145/3491102.3501825.

Loftsdottir, Dagmar, and Matthew Guzdial. 2022. “SketchBetween: Video-to-Video Synthesis for Sprite Animation via Sketches.” In *Proceedings of the 17th International Conference on the Foundations of Digital Games*. A.No. 32: 1–7. https://doi.org/10.1145/3555858.3555928.

Mital, Parag K, Mick Grierson, and Tim J Smith. 2013. “Corpus-based visual synthesis: an approach for artistic stylization.” In *Proceedings of the ACM Symposium on Applied Perception*, 51-58. https://doi.org/10.1145/2492494.2492505.

Mitchell, Melanie. 2006. “Complex systems: Network thinking”, *Artificial intelligence*, 170: 1194-212. https://doi.org/10.1016/j.artint.2006.10.002.

Navas, Eduardo. 2022. *The Rise of Metacreativity: AI Aesthetics After Remix* (Taylor & Francis). isbn: 9781003164401.

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 preprint arXiv*:2112.10741.

Popov, Vadim, Stanislav Kamenev, Mikhail A Kudinov, Sergey Repyevsky, Tasnima Sadekova, Vitalii Bushaev, Vladimir Kryzhanovskiy, and Denis Parkhomenko. 2020. “Fast and Lightweight On-Device TTS with Tacotron2 and LPCNet.” In I*NTERSPEECH*, 220-24.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021. “Learning transferable visual models from natural language supervision.” In *International conference on machine learning*, 139:8748-8763. PMLR.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language models are unsupervised multitask learner “, *OpenAI blog*, 1: 9.

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: 1623-37.

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.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684-95.

Saharia, Chitwan, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. “Palette: Image-to-Image Diffusion Models.” In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, 1-10.

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.

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.” In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4779-83. IEEE.

Steyerl, Hito. 2019. “This is the Future”, Accessed May 11, 2023. https://yaci-international.com/hito-steyerl-this-is-the-future-2019/.

Whittington, James CR, Joseph Warren, and Timothy EJ Behrens. 2021. “Relating transformers to models and neural representations of the hippocampal formation”, *arXiv preprint arXiv*:2112.04035.

Yang, Ling, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. “Diffusion models: A comprehensive survey of methods and applications”, *arXiv preprint arXiv*:2209.00796.

Ye, Zhiyuan, Chenqi Xue, and Yun Lin. 2021. “Visual perception based on gestalt theory.” In *International Conference on Intelligent Human Systems Integration*, 792-97. Springer. doi: https://doi.org/10.1007/978-3-030-68017-6_118.