Article draft to appear in Springer (2025).

14th International Conference EvoMUSART 2025, EvoStar 2025, Trieste, Italy, April 23–25, 2025, Proceedings.

Arandas, L., Ionescu, I., Khan, M., Grierson, M., Carvalhais, M. (2025). All YIN No YANG: Geometric Abstraction of Oil Paintings with Trained Models, Noise and Self-reference. In: Machado, P., Johnson, C., Santos, I. (eds) Artificial Intelligence in Music, Sound, Art and Design. EvoMUSART 2025. Lecture Notes in Computer Science, vol 15611. Springer, Cham. https://doi.org/10.1007/978-3-031-90167-6_17.

All YIN No YANG: Geometric abstraction of oil paintings with trained models, noise and self-reference

Luís Arandas[1], Iulia Ionescu[1], Murad Khan[1],
Mick Grierson[1], Miguel Carvalhais[2]
[1] University of the Arts London, Creative Computing Institute, London, UK
[2] i2ADS, Faculty of Fine Arts University of Porto, Portugal
{l.pintoarandas, i.ionescu, murad.khan, m.grierson}@arts.ac.uk, m.carvalhais@fba.up.pt

Abstract. The rapid development of Diffusion models and the declarative nature of interfaces developed for the public require automation methods, where media production can harness natural language as a mode of representation but not necessarily of interaction with humans. This article describes an image-to-video Diffusion system which removes practitioners from the process of defining prompts when producing images with conditional reference, documenting a set of results with a custom dataset of oil paintings. Our research focuses on the appropriation of trained model ensembles that are coordinated to produce indefinite sets of frames with occasional human intervention utilising timeline-based architectures. The proposed system automates a CLIP-guided DDPM with a supplementary depth estimation model and through a set of compositing techniques we found that results with coincidental and diverging descriptions can be useful for moving-image element composition. Our experiments focus on the representation of human figure and its morphological transformation.

Keywords: Language-guided automation, Oil painting diffusion, Predictive abstraction, Depth-mask compositing, Timeline-based architectures.

1 Introduction

Over the last years there has been a shift in the world of computational image generation. Though practitioners have long experimented with generative Deep Learning (DL) models to produce digital media, the development of text-guided adversarial nets (GAN) and Diffusion models such as DALL-E and Stable Diffusion shifted the boundaries of still and moving image practice [34]. The rise of networks such as Contrastive Language-Image Pre-Training (CLIP) [25] drove a shift in the use of multimodal network architectures where natural language describes archives, photographs and paintings [8]. Multimodality appears in generative practice as a form of relationship between data spaces [3], regarding the internal architecture of models capable of capturing patterns and maintaining or updating them as a memory system [10].

Using dual representations in the same Neural Network (NN), generative DL methods allow creators to produce work on representation, abstraction and divergence projecting new personal data into models and exploring the relationships conceived through scores and embeddings [17]. Despite the initial interest in the use of CLIP to condition GANs, the shape shift to Diffusion models was indicative of a movement in the field towards new architectures for practitioners. In particular, the use of natural language "prompts" to steer the synthesis process has become a staple feature of generative practice over the past five years and has given rise to the methods of "prompt engineering", "prompt-chaining" and even "prompt-based LLM reasoning" [20, 22]. Prompt engineering can be best understood as the construction of syntactically and semantically specific strings of text that can be supplied to a model to condition its output [12]. Whilst both still and moving image works have been produced using prompt-based Diffusion models [33] there remains a growing need to articulate the limits of these practices and to experiment with the use of natural language within the generative pipeline outside of the input and declarative stage.

This article describes a system working as an extension to CLIP-guided DDPM pipelines and we document video (outputs) produced using as custom dataset of oil paintings (inputs), developed within the collaboration *All YIN No YANG*. Our approach to multimodal image synthesis focuses on the removal and reimagination of the human figure, visually abstracting still image inputs, defining prompts as descriptions, and extending image frame

computation with a separate depth estimation model for compositing.[1] We document a working timeline-based image-to-video Diffusion architecture and dilute human influence on the prompting process. Our research promotes language-guided DDPMs as automatic mechanisms for creative practice and possibly part of the underlying structure of co-creative agents, where the representation of language works independent of human specification.

2 Diffusion and reversing from noise

Image diffusion models represent a different computing paradigm of generation to GANs but ultimately seek to: capture and understand a data distribution such that we can create novel or unseen samples that approximate the original distribution. Diffusion probabilistic models were proposed as a parameterised Markov chain trained using variational inference; models that systematically destroy and reconstruct structure in a given data distribution through a two-part process: (1) An iterative *forward* process in which Gaussian noise (with variance) is applied stepwise to an input sample until is corrupted and (2) A *reverse* process of learning to a desired distribution, denoising samples and restoring structure to the perturbed input [30]. The *reverse* process is usually modelled using NNs for the complexity trying to approximate, [13] showed an equivalence between Denoising Diffusion Probabilistic Models (DDPM) and score-based generative models, and generate high quality images; see [9] specific architecture used throughout this research. In conditional diffusion of images from text, researchers found great success in conditioning UNets in the diffusion model framework in both time and text, where the network is responsible for mapping the signal to its noise [27].

Diffusion processes of approximation have been used extensively in still and moving image also with Cascaded Diffusion Models (CDM) which are pipelines of independent diffusion models that generate images of increasing resolution [14, 29]. *Forward-reverse* diffusion can generate material around a reference input or directly from noise, absent the use of language, and predict frame sequences with an anchor to some arbitrary space defined at a past period of learning [11]. The addition of recursivity and of autonomous behaviour using CLIP at each step, can be used to aid the composition of new systems that produce movement across frames by pointing inwards and to fragments of themselves; models which crystallised some visual world in numerical space [6, 7]. From noise specific shapes can be found, as if pixels were contracting in time with a direction, like a simulation of spray particles on paper, being the process of *forward-reverse* a play between pixels oscillating from low to high concentration in 2D or 3D space. Diffusion is commonly used in image processing such as to fill gaps in data distributions, denoising, and out-painting, these are practical implementations of this process within the industry [4]. On a practical case, diffusion with classifier guidance allows to generate image frames from text prompts, where a trained diffusion model score estimate can be computed with the gradient of a separate image classifier [9], also explored with latent models, with visual definitions based on specific labels or clusters of data in each trained representations [26].

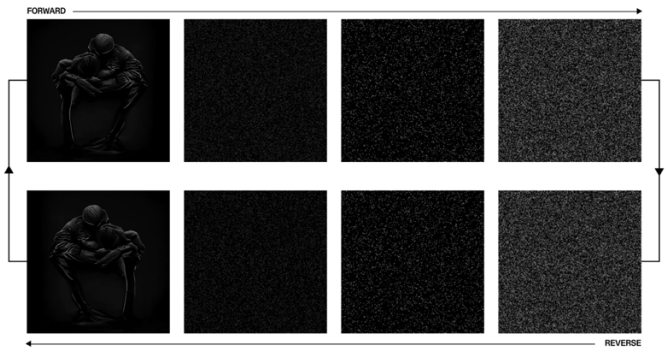

Fig. 1. Diffusion *forward-reverse* (top-down) process illustration. Formal variation on a specific input frame by computing a noise scheduler and learning to reverse back.

[1] Without explicit class or prompt definition, Depth Estimation models proved able to identify human figures within backgrounds, see 31. Su, P.-C. and M.-T. Yang, *Integrating Depth-Based and Deep Learning Techniques for Real-Time Video Matting without Green Screens.* Electronics, 2024. **13**(16): p. 3182.

3 Conditional guidance and natural language

Classifier-free guidance [15] as found in GLIDE, shows guidance can be performed by a pure generative model without a separate classifier. The trade-off between diversity and fidelity is up to the practitioners and both DALL-E 2, Imagen and conditional diffusion system *Disco-Diffusion* [8], allow to use natural language to guide the synthesis process [33]. These models and architectures guide the distribution prediction towards generating an image with greatest visual similarity to the prompt [22]. In the case of moving image, text-guided diffusion can carry out this process over an additional temporal domain, presenting additional opportunities compared to still image generation (e.g., sets of prompts, control on variance and image-to-image correspondence) [18]. We start by illustrating the CLIP-guided diffusion process we build upon:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t\, \mathbf{I}\right)$$

In order to produce DDPM, the system uses a UNet to map noisy frames to clean frames, conditioned on textual input. In the previous equation we illustrate the *forward* (diffusion) process where β_t is noise schedule. The *reverse* process aims to recover clean frames by approximating $p_\theta(x_{t-1}|x_t)$ using a NN parameterised by θ. The UNet predicts the target noise $\epsilon_\theta(x_t, t)$ estimating a clean frame:

$$x_0 = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \sqrt{1-\overline{\alpha}_t}\,\epsilon_\theta(x_t, t)\right)$$

The CLIP model guides the diffusion process by computing a similarity loss between the generated frames and text embeddings, ensuring semantic alignment. To extend this process over N frames with variance and create a chain of *forward-reverse* where each image influences the next, we modify the reserve step for every frame f by incorporating a blending of the previous clean frame x_0^{f-1} into the estimation of the current clean frame x_0^f (time correspondence, high or low variability across frames). We illustrate this process achieved using a weighting factor γ controlling the influence of previous frame:

$$x_0^f = (1-\gamma)\left(\frac{1}{\sqrt{\overline{\alpha}_t^f}}\left(x_t^f - \sqrt{1-\overline{\alpha}_t^f}\,\epsilon_\theta\left(x_t^f, t\right)\right)\right) + \gamma\, x_0^{f-1}$$

3.1 Description of model architecture

Following the previously stated objectives of research, we enumerate some design guidelines: 1) To produce a system which runs independently after image input delivering finalised video sequences, and 2) explore quantifiable metrics for divergence given the set of trained models with time-based geometrical and semantical conflicts. Having a dataset of images with scanned oil paintings we describe both a DDPM and a feedback loop adding a Depth estimation MiDaS model, allowing to produce sets of frames [24]. By producing renders with *frame:prompt* correspondence over arbitrary lengths we also resort to timeline objects, see a timeline-based coordination system for trained AI models and feedback equilibrium in short-film length (present-future frame correspondence) in [1].

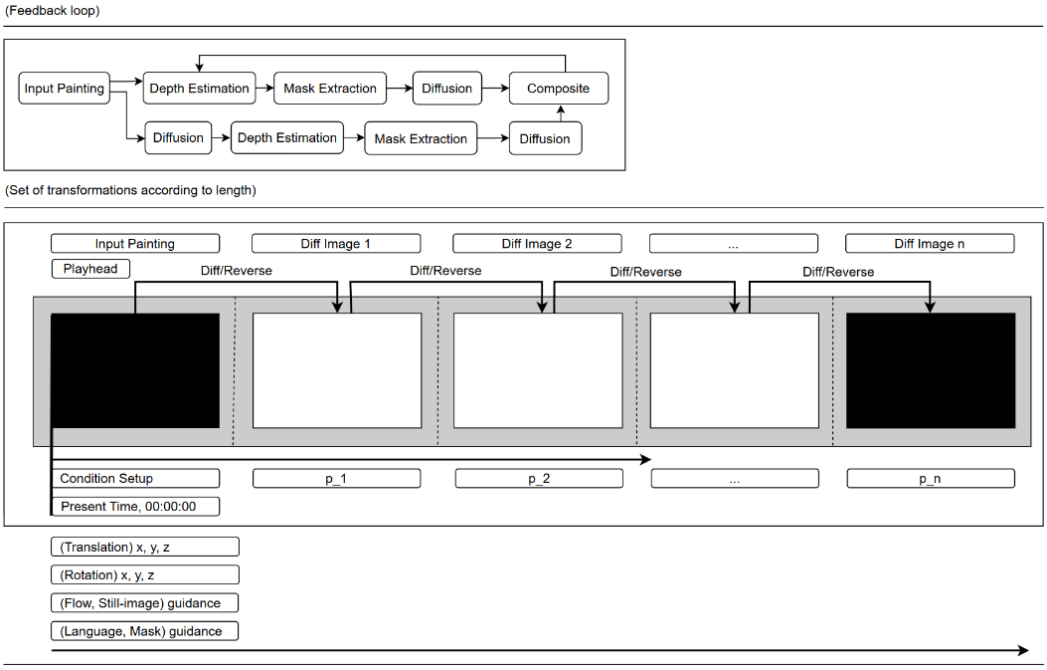

Fig. 2. Illustration of the diffusion processing flow with added modules.

If we consider the previous DDPM and Figure 2, the feedback loop produces collages that work on the initial painting background over arbitrary periods of length. To illustrate, we compute a depth map D^{f-1} of the image x_0^{f-1}, and extract alpha masks M^f from D^{f-1}; $D^{f-1} = \text{DepthEstimator}(x_0^{f-1}), M^f = \text{MaskExtractor}(D^{f-1})$. We then apply conditional diffusion on the noisy frame x_t^f conditioned on masks M^f and prompt p:

$$\tilde{x}_0^f = \frac{1}{\sqrt{\overline{\alpha}_t^f}}\left(x_t^f - \sqrt{1-\overline{\alpha}_t^f}\,\epsilon_\theta\left(x_t^f, t, M^f, p\right)\right)$$

Where: 1) $\tilde{x}_0^f$ is the intermediate clean frame estimate at frame f; 2) x_t^f the noisy frame at frame f and time step t; 3) $\overline{\alpha}_t^f = \prod_{s=1}^{t} \alpha_s^f$ is the cumulative product of $\alpha_t^f = 1 - \beta_t^f$. Next we combine the intermediate frame with the original frame (oil painting) using the mask, resulting in the compositing step:

$$x_0^f = M^f \odot \tilde{x}_0^f + (1 - M^f) \odot x_0^{f-1}$$

The previous equation illustrates: 1) M^f defines the regions where the new content $\tilde{x}_0^f$ should be applied; 2) $(1 - M^f)$ specifies the complementary regions to retain from the original frame x_0^{f-1}; $\odot$ denotes element-wise multiplication. This compositing process integrates the newly generated content $\tilde{x}_0^f$ into the regions specified by mask M^f, retaining the background from the previous frame x_0^{f-1}. In the feedback loop, we intentionally diffuse the foreground mask before extracting masks, keeping the original oil painting as the background. The workflow adds several steps to the *forward-reverse* CLIP-guided DDPM: 1) Computing depth at every diffused image; 2) Use the resulting mask M^f to isolate the foreground; 3) Apply conditional diffusion inside the masked region; 4) Compute depth on the final diffusion and use this as the foreground in subsequent steps. This process enables the system to decide where to draw based on the depth masks of each painting, allowing features such as faces or torsos to guide the placement of new images, with automation on the process of natural language description.

4 All YIN No YANG

All YIN No YANG is a research collaboration approaching the question of Human-AI interaction through a practice-led inquiry into the varieties of formal and aesthetic divergence within images. This practice-based collaboration focuses on exploring contemporary methods for image generation from noise as a method of reverse by self-reference, using trained models that learn and are understood from the standpoint of translation (image-to-video and sequence-to-sequence), divergence or even abstraction [2, 5]. Diffusion models and DDPMs promote a space for experimentation, where still images can be manipulated through AI representations to create moving images, while searching for new computing methods that reveal the process in which we base our practice. Within this research we focus specifically on the step-based Diffusion pass of using both a custom CLIP-guided DDPM and a custom dataset of oil paintings.

Forward-reverse processes of Diffusion for image generation can be dissected into parts, namely steps which can be skipped or interpolated with scores from adjacent models, making video outputs a combination of complementary predictions, where each trained model contributes to a bigger architecture that abstracts more tractable editing and sequencing mechanics. The techniques we build upon have been applied to multiple fields of experimental video art and film previous to this research, and we place our results within the conception of the human body from simple materials. Our work is resistant to the idea of beauty being necessarily tied to failure and success within Creative and Generative AI outputs and as further explained in the following sections, is precisely where we can find new artefacts or methods for video composition (e.g. coincidental and diverging prompts with the same image condition).

Fig. 3. Initial render: (Left) Input painting; (Right) Snapshot of video output. Illustrates image-to-video CLIP-guided DDPM without depth model addition and feedback loop with constant variance.

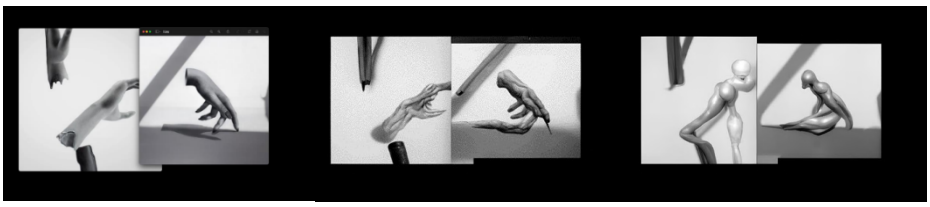

Fig. 4. Illustration of image-to-video CLIP-guided DDPM without depth model addition and feedback loop with random variance and text prompt, input paintings into different materials.

4.1 Materials and results

We started defining two variation targets (two paintings, Fig. 5), each with white backgrounds and preview frames, implementing the defined image-to-video CLIP-guided DDPM system with the additional depth model; see [23] for an opposite implementation using YOLO network. We document the results with coincidental and diverging prompts (derived from Target 1 – Left, or Target 2 – Right), see next Figure for visual reference: Description 1) “Two black and white paintings of a woman’s face.”, and Description 2) “A black and white photo of a bird and woman.”, automatically derived [32]. We produce two videos both with the same background Target 1, with natural language description of Target 1 and Target 2, Figure 6 illustrates several video snapshots.

Fig. 5. Arrangement of target oil painting scans, keeps static with zero movement for the duration of each output video. (Left): Background for output video sequences (Description 1), (Right): Divergent source of natural language (Description 2).

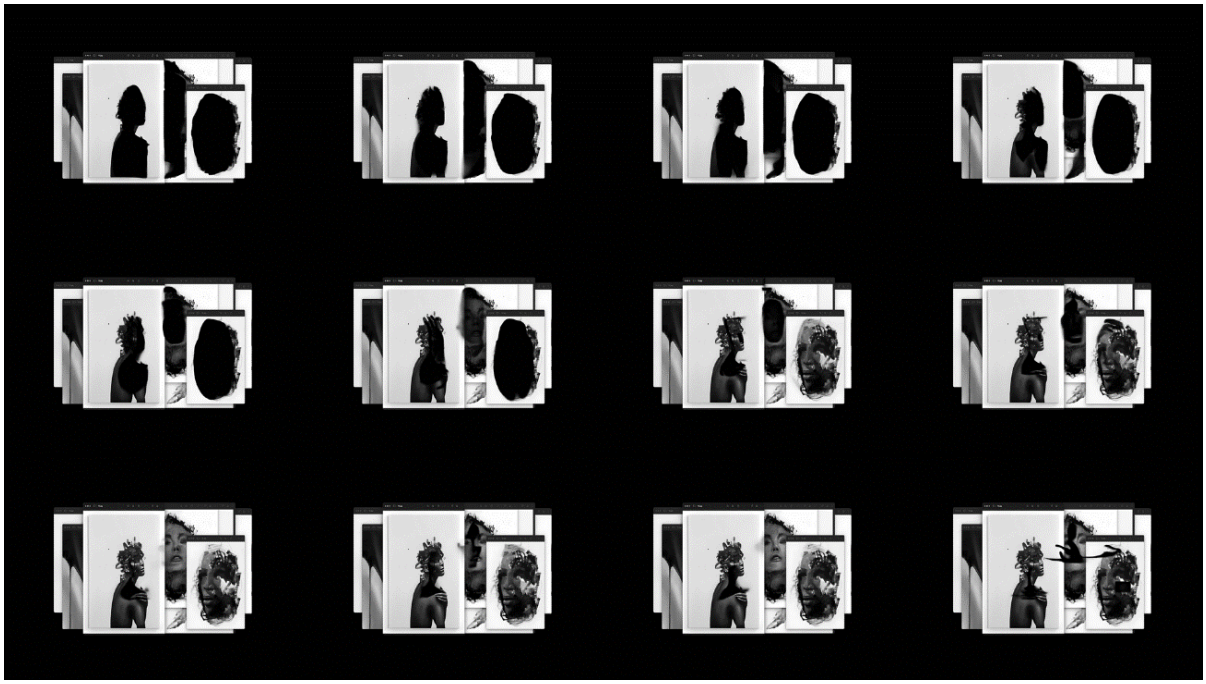

Fig. 6. Output videos: set of 4 strips demonstrating mask diffusion across frames with constant background illustrating: coincidental and diverging guidance, random variance and application of compositing (described in 3.1). Further 2 snapshots of the 2 resulting videos with text guidance as explained in section 4.1.

4.2 Analysis and proposal

The proposed depth-based extension to the described image-to-video DDPM system proved to be useful when compositing diffused images over an arbitrary background input, demonstrating morphological variations of the oil paintings. The experiments conducted show that diffusing inside depth masks with non-constant variance can produce experimental (as related to image-frame *structure*) but valuable results, that by the nature of compositing are a (video) collage of a variation into the original reference. Without class pre-definition on the analysis of the oil paintings we resort to the AI depth estimation model consistency, which proves useful in our dataset, e.g.,

object positioning. We design the system on top of known CLIP-guided DDPM processes and show results with semantic consistency over the inputs. The proposed design can be set to real-time and benefit from adjacent controls regarding, e.g., object detection and 3D representation [19, 21, 28].[2]

5 Conclusion

DDPM-based systems design can benefit from automation, where model ensembles can be coordinated to produce videos from arbitrary still image inputs. Semantic specificity can easily render models as tools rather than co-creative agents, this research extends current methods for video Diffusion developing an automated system to visually represent the human across frames; as a variation of the input space. We propose a CLIP-guided DDPM image-to-video pipeline where a complementary depth estimation model extends a timeline-based architecture for compositing. Our experiments focus on rendering abstractive videos (outputs) from oil painting scans (inputs) with reduced human intervention identifying potential real-world use-cases. As a broader contribution, the collaboration *All YIN No YANG* represents an attempt to explore Diffusion-based image-making and its impact upon the development of the artist and their practice, where DL tools are used beyond a purely instrumental view and situated as co-constitutive agents in creative practice.

Acknowledgments: Research leading to these results was initially 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 with residency at Camberwell College of Arts, and by UAL Creative Computing Institute Research and Knowledge Exchange.

Disclosure of Interests: The authors declare no competing interests.

References

1. Arandas, L., M. Grierson, and M. Carvalhais, *Computing Short Films using Language-guided Diffusion and Vocoding through Virtual Timelines of Summaries.* Insam Journal, 2023. **10**.
2. Arandas, L., et al. *all YIN no YANG: Automating Language-guided Diffusion Systems in Search of Abstraction*. in *Explorations on Sound and New Media Art Conference, book of abstracts*. 2023.
3. Baltrusaitis, T., C. Ahuja, and L.P. Morency, *Multimodal Machine Learning: A Survey and Taxonomy.* IEEE Trans Pattern Anal Mach Intell, 2019. **41**(2): p. 423-443.
4. Blattmann, A., et al., *Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models.* arXiv preprint arXiv:2304.08818, 2023.
5. Candy, L., *Practice based research: A guide.* CCS report, 2006. **1**(2): p. 1-19.
6. Crawford, K. and T. Paglen, *Excavating AI: the politics of images in machine learning training sets.* Ai & Society, 2021.
7. Croitoru, F.-A., et al., *Diffusion models in vision: A survey.* arXiv preprint arXiv:2209.04747, 2022.
8. Crowson, K., et al., *VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance.* 2022.
9. Dhariwal, P. and A. Nichol, *Diffusion models beat gans on image synthesis.* Advances in Neural Information Processing Systems, 2021. **34**: p. 8780-8794.
10. Elgendy, M., *Deep learning for vision systems.* 2020.
11. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning.* MIT Press, 2016.
12. Gu, J., et al., *A systematic survey of prompt engineering on vision-language foundation models.* arXiv preprint arXiv:2307.12980, 2023.
13. Ho, J., A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models.* Advances in neural information processing systems, 2020. **33**: p. 6840-6851.
14. Ho, J., et al., *Cascaded diffusion models for high fidelity image generation.* Journal of Machine Learning Research, 2022. **23**(47): p. 1-33.

[2] Numerical transformations for warp methods inherit from NVIDIA's 16. Ilg, E., et al. *Flownet 2.0: Evolution of optical flow estimation with deep networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

15. Ho, J. and T. Salimans, *Classifier-free diffusion guidance.* arXiv preprint arXiv:2207.12598, 2022.
16. Ilg, E., et al. *Flownet 2.0: Evolution of optical flow estimation with deep networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
17. Khandelwal, A., et al. *Simple but effective: Clip embeddings for embodied ai*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
18. Kim, G., T. Kwon, and J.C. Ye. *DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
19. Li, L. *The Impact of Artificial Intelligence Painting on Contemporary Art From Disco Diffusion's Painting Creation Experiment*. in *2022 International Conference on Frontiers of Artificial Intelligence and Machine Learning (FAIML)*. 2022. IEEE.
20. Li, Y., R. Zhang, and J. Liu. *An enhanced prompt-based LLM reasoning scheme via knowledge graph-integrated collaboration*. in *International Conference on Artificial Neural Networks*. 2024. Springer.
21. Liu, S., et al. *Adaattn: Revisit attention mechanism in arbitrary neural style transfer*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
22. Liu, V. and L.B. Chilton, *Design Guidelines for Prompt Engineering Text-to-Image Generative Models*, in *CHI Conference on Human Factors in Computing Systems*. 2022. p. 1-23.
23. Liu, Y., H. Zhang, and D. Gao, *DiffYOLO: Object Detection for Anti-Noise via YOLO and Diffusion Models.* arXiv preprint arXiv:2401.01659, 2024.
24. Padkan, N., et al., *Evaluating Monocular Depth Estimation Methods.* International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2023. **48**(1): p. 137-144.
25. Radford, A., et al. *Learning transferable visual models from natural language supervision*. in *International conference on machine learning*. 2021. PMLR.
26. Ramesh, A., et al., *Hierarchical text-conditional image generation with clip latents.* arXiv preprint arXiv:2204.06125, 2022.
27. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. 2015. Springer.
28. Rosinol, A., J.J. Leonard, and L. Carlone. *Nerf-slam: Real-time dense monocular slam with neural radiance fields*. in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2023. IEEE.
29. Saharia, C., et al., *Palette: Image-to-Image Diffusion Models*, in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*. 2022. p. 1-10.
30. Sohl-Dickstein, J., et al. *Deep unsupervised learning using nonequilibrium thermodynamics*. in *International Conference on Machine Learning*. 2015. PMLR.
31. Su, P.-C. and M.-T. Yang, *Integrating Depth-Based and Deep Learning Techniques for Real-Time Video Matting without Green Screens.* Electronics, 2024. **13**(16): p. 3182.
32. Wu, K., et al. *Tinyclip: Clip distillation via affinity mimicking and weight inheritance*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
33. Yang, L., et al., *Diffusion models: A comprehensive survey of methods and applications.* arXiv preprint arXiv:2209.00796, 2022.
34. Zhang, C., et al., *Text-to-image diffusion model in generative ai: A survey.* arXiv preprint arXiv:2303.07909, 2023.