EpicDiffusion: Enhancing video diffusion models for storytelling

Group 7

Abhiram Vadlapatla
ASU ID: 1225050671
avadlap1@asu.edu

Sri Valli Kanumuri
ASU ID: 1231607171
skanumu5@asu.edu

Manikandan Sundararaman
ASU ID: 1231923786
msunda17@asu.edu

Dhruv Bindra
ASU ID: 1229592899
dbindra@asu.edu

Kaustubh Uday Kulkarni
ASU ID: 1229592483
kukulkar@asu.edu

Abstract

Diffusion models are good at generating images conditioned from textual prompts. Recent advancements in encoding video information in diffusion models have led to the generation of temporally consistent videos, using **video diffusion models**. Yet, there is a lot of exploration left in the domain of translating passages from novel and epic literature to animated illustrations. This project aims to improve diffusion models so that they can better generate animated illustrations from passages in novels. We discuss the problems in diffusion models, LLM integration for generating prompts, and a custom architecture to support text-to-video alignment. We introduce a new LLM guided prompt weighting method to generate spatially and temporally consistent videos for fictional passages. Furthermore, we perform qualitative and quantitative evaluations on these images and videos across various action categories.

1 Problem Statement

The objective of EpicDiffusion is to address the current limitations of video diffusion models in capturing the nuanced narrative and thematic elements essential for epic and historic fiction. Despite their success in a broad range of tasks, these models often fall short in generating contextually rich visuals for specific storytelling genres, resulting in outputs that may lack depth or relevance. EpicDiffusion seeks to develop a bespoke video generation model that:

- Enhances the generative capabilities of diffusion models to accurately interpret and visualize complex narrative content.
- Advances the technology while respecting the source material's artistic and historical integrity, providing a tool for generating personalized, engaging visuals.

1.1 Motivation

The motivation behind EpicDiffusion is twofold:

- **Enhanced Reading Engagement**: Integration of visual elements into epic and historic fiction can significantly enrich the reader's experience. By generating dynamic visuals,

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

EpicDiffusion aims to bridge traditional reading experiences with contemporary digital storytelling methods.

- **Personalized Video Graphics**: Tailoring visuals for book covers and illustrations to individual preferences or specific narrative elements can transform reading into an interactive experience, enhancing enjoyment and emotional connection to the story.

2 Technical Background

2.1 Diffusion Models

A diffusion model is a generative model, which means it generates data based on what it is trained on.

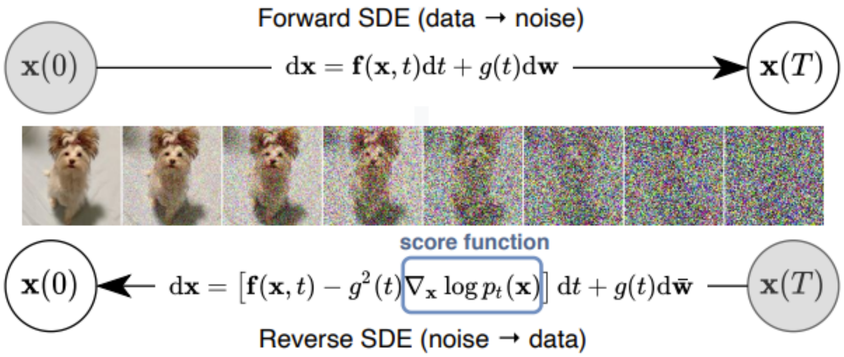

Forward SDE (data → noise)
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
$\mathbf{x}(T)$
score function
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
$\mathbf{x}(T)$
Reverse SDE (noise → data)

Figure 1: Noising and denoising process in a diffusion model.

Diffusion models involve three processes:

1. **Forward Process (Diffusion or Noising)**: Beginning with input data such as an image, the model adds Gaussian noise sequentially, until the data becomes noisy. Adding noise is guided by a Markov chain, which dictates how the noise is added at every step.
2. **Reverse Process (Denoising)**: Once the data is noisy, the model attempts to reverse the process and fetch the original image. Through this process, the model learns to predict the reverse steps required to reconstruct original data from the noisy one.
3. **Generation**: Once the model is trained through progressive noising and denoising, it generates new data starting with random noise and creates the data based on what it perceived during training.

The key to diffusion models is to carefully balance the noising and denoising processes to make the model learn complex data distributions.

The general architecture for an image generation latent diffusion model can be summarized as follows: Here the diffusion process happens in the latent space where the image information is encoded in lower dimensions. The architecture primarily consists of three components:

- **Autoencoder**: Used to encode the image information into a latent representation
- **Text Encoder**: Used to convert the prompt text while inference or text in image-text pair while training to embeddings
- **U-Net**: U-Net is primarily used for the noising/denoising process. If the model is conditioned based on text/other modalities, it contains cross-attention layers to assess the alignment of the conditional input(text) to the output image

In conclusion, we understand that improving the prompt or improving the latent space(UNet) are two major target areas to enhance diffusion models. We explore related work in this direction below.

2.2 Prompt weighting

Prompt weighting is a technique used in generative models to control the influence of different concepts within a text prompt on the generated image. It involves adjusting the scale of the text embeddings corresponding to each concept in the prompt to emphasize or de-emphasize them during image generation. This control allows for more precise guidance of the model's attention, leading to better-quality outputs.

Overall, prompt weighting offers flexibility and control in tailoring the generated images to specific desired concepts or themes.

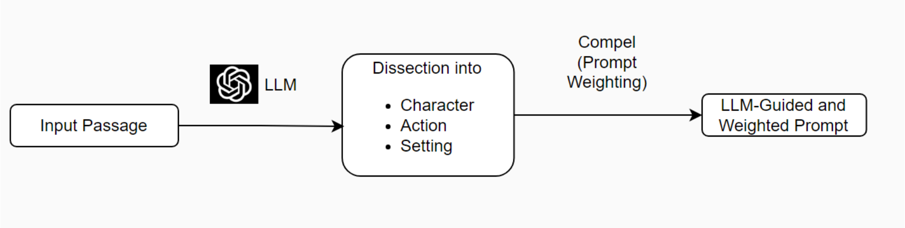


Figure 2: Prompt Weighting

2.3 FreeU

FreeU is a method to enhance the generation quality of image and video models, particularly those based on the diffusion U-Net architecture. It strategically re-weighs the contributions sourced from the U-Net's skip connections and backbone feature maps, to leverage the strengths of both components of the U-Net architecture

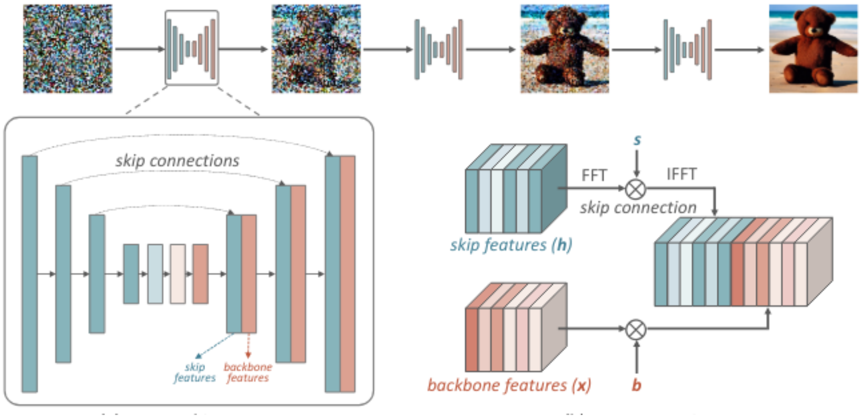


Figure 3: FreeU Architecture

- **UNet**: The U-Net architecture sometimes overlooks the semantic information provided by the backbone of the network, focusing too much on introducing high-frequency details. To address this issue, FreeU strategically adjusts the contributions from both the skip connections and the backbone feature maps of the U-Net.

By re-weighting these contributions, FreeU leverages the strengths of both components, improving the overall generation quality. Importantly, this enhancement is achieved without the need for additional training or fine-tuning, making it a convenient and efficient solution.

3 Related Work

3.1 I2VGenXL Diffusion Model

I2VGen-XL is a video-synthesis approach designed to address challenges in semantic accuracy, clarity, and spatiotemporal continuity in video generation. It operates in two main stages:

Base Stage: Ensures semantic coherence and content preservation from input images, using two hierarchical encoders.
Refinement Stage: Focuses on enhancing video resolution to 1280x720 and improves details in video through additional brief text inputs.

The model's performance is enhanced by a vast dataset comprising 35 million text-video pairs and 6 billion text-image pairs, increasing diversity and quality in the generated videos. I2VGen-XL stands out for its ability to decouple semantic and qualitative factors, using static images to guide the alignment of input data, resulting in videos with improved semantic accuracy, detailed continuity, and clarity.

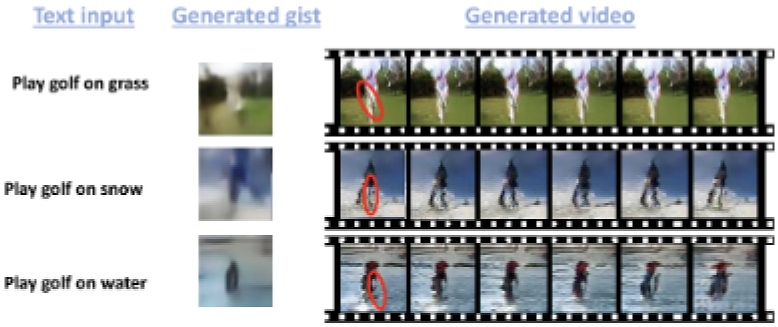


Figure 4: I2VGenXL video diffusion model

This model is used in our image to video generation process and it's action prompt is guided by an LLM.

3.2 LLM-Grounded Diffusion

LLM-Grounded Diffusion is a prompt-enhancing method that guides text-to-image diffusion models to handle complex prompts involving numeracy and spatial reasoning. It uses a two-stage process with a pretrained Large Language Model involving the following:

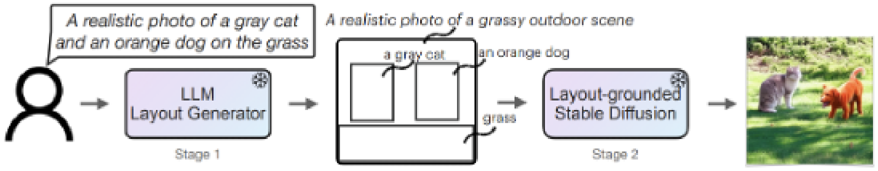

A realistic photo of a gray cat and an orange dog on the grass
LLM Layout Generator
Stage 1
A realistic photo of a grassy outdoor scene
a gray cat
an orange dog
grass
Layout-grounded Stable Diffusion
Stage 2

Figure 5: LLM Grounded Diffusion

- Creating a scene layout with captioned bounding boxes.
- A controller-guided diffusion model to generate images based on the scene layout.

This approach improves generation accuracy, supports multi-round scene specification, and works with prompts in various languages, all without needing extra model training. It aims to boost creativity by more accurately interpreting complex prompts.

3.3 Timesformer

Video understanding models take a video as input and return the key action performed in the video. Recent advancements in attention and image transformers have led to the development of self-attention and cross-attention mechanisms across frames in videos, which highly improve action classification tasks.

Recent state of the art model in Video Understanding is Timesformer from Facebook. It introduces a spatial and temporal attention mechanism and provides 70% in action recognition tasks.

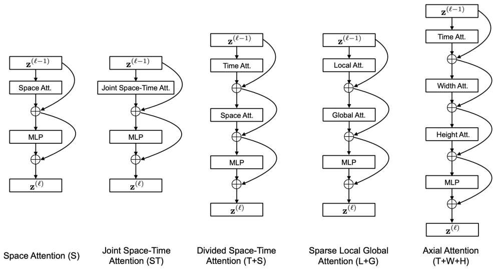


Figure 6: Attention in timesformer model

3.4 Kinetics 400 dataset

This dataset is the standard in video understanding tasks. It contains around 650k videos across 400 categories. It is primarily used to train video understanding tasks, and the labels can help us identify temporal consistency with generated videos.

4 Technical Implementation

4.1 Infrastructure

For executing EpicDiffusion, we were utilizing ASU Sol Supercomputing resources, particularly the Nvidia A100 GPU, to generate the Stable Diffusion image and the video. For referencing the model architectures and implementation, we have used HuggingFace's Stable Diffusion and Image-to-Video-Gen-XL model.

4.2 Datasets

To test the efficiency of our approach, we used existing fictional passages from fictional content such as Aesop's fables and Adventures of Tom Sawyer. For evaluation with CLIP score and Video understanding models, we generated synthetic data across different action categories. We picked the action categories from the state of art dataset called Kinetics-400. This choice was made since the video understanding models are trained on the Kinetics 400 dataset and predict the action in the video accordingly.

4.3 Architecture

- Firstly, we prompt the input passage to an LLM(GPT-4 in this case) and ask it to dissect the given passage into sections such as Character, Action, and Setting. This is done using the Chain-of-thought (CoT) method to enhance the quality of analysis.
- Using this list, we again use chain-of-thought and prompt the LLM to generate an effective prompt with weights. For example, if the prompt "a cat and a red ball" needs to focus on both the cat and red ball, we introduce a weight as follows "a (cat)1.5 and a (red ball)1.5" to give equal weightage to both the items.
- We configure the model using Guiding-Scale and Inference-Steps parameters of the text-to-image and image-to-video models to enhance the generation capabilities of the model.

- Along with these parameters, we also add a FreeU Mixin to the diffusion pipeline, where we can tweak the skip connection contributions from the U-Net. This helps us to rebalance the negative outputs from the UNet as we increase the denoising steps.
- Using the generated image, and the Action text from the previously dissected prompt, we pass it to the I2VGenXL model, generating the short video based on the image and the action as the guided text.

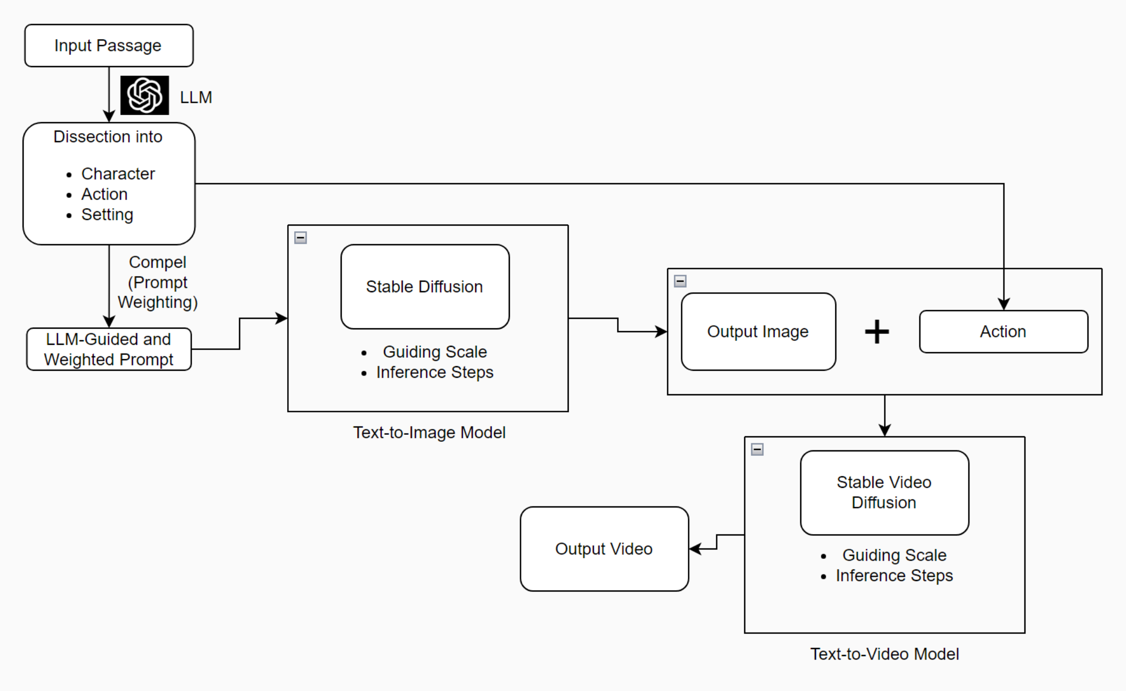


Figure 7: EpicDiffusion architecture

5 Evaluation plan

5.1 Evaluating T2I model

- Fantasy Fiction: For this category, we evaluated the outputs qualitatively with manual inspection, since fictional attributes are tough to capture via any existing benchmarking methods like CLIP Score.
- Historical Fiction: Similar to the fantasy fiction category, we have a problem with existing benchmarking methods to assess historical attributes in an image. Hence, a qualitative manual inspection method was used.
- Contemporary: For contemporary fiction-based passages, we picked passages that confirm to actions from the Kinetics 400 dataset. We used those passages to calculate the CLIP score according to the ground truth label from the dataset. Overall, across 3 categories for 10 images, our method generates an average CLIP score of 0.26 whereas the default prompting method generates an average CLIP score of 0.21

5.2 Evaluating I2V Model

Evaluating Video Generation is a tedious task since it needs to assess text alignment across frames. For using approaches like FVD, we require real videos for reference. Since our content is generated from fictional content, and the I2V model is primarily used to condition an action from a generated image, we utilized Timesformer to compute the action classes from 30 videos across 3 classes. We then compute the overall accuracy from the ground truth. This process is depicted in Figure 9.

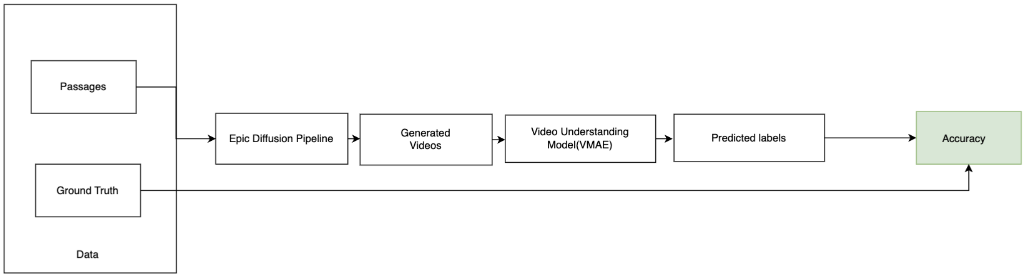


Figure 8: Evaluation pipeline for Video Analysis

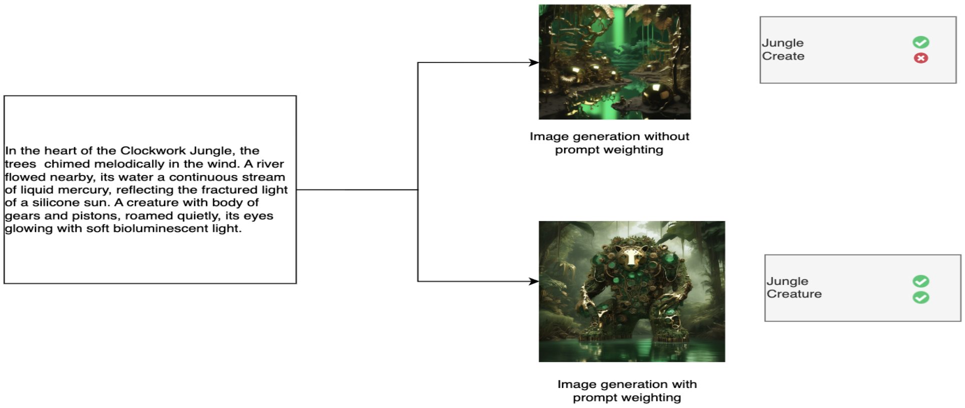

In the heart of the Clockwork Jungle, the trees chimed melodically in the wind. A river flowed nearby, its water a continuous stream of liquid mercury, reflecting the fractured light of a silicone sun. A creature with body of gears and pistons, roamed quietly, its eyes glowing with soft bioluminescent light.

Image generation without prompt weighting
Jungle ✓
Create ✗
Image generation with prompt weighting
Jungle ✓
Creature ✓

Figure 9: Passage-Image alignment for Fantasy Fiction

6 Results

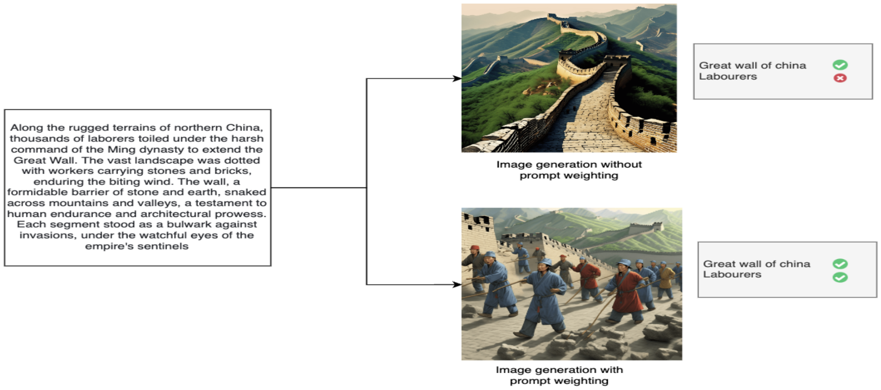
6.1 Semantic composability of T2I model

Fantasy Fiction: In Figure 9, image diffusion model generates a highly realistic image for fictional forest with river of mercury, but entirely ignores the creature in the next sentence. With prompt weighting from LLM, the image focuses on creature as well and generates an image with both the forest and the creature
Historical Fiction: In Figure 10, image diffusion model generates a highly realistic image for Great wall of china, but entirely ignores the labourers and the historical context. With prompt weighting from LLM, the image focuses on labourers and generates an image with both the labourers and the Great wall of China **Contemporary Fiction**: In Figure 11, image diffusion model generates a highly realistic image of a man in a car. But ignores the fact that the person is standing in traffic. The second image with prompt weighting takes into the account the context of the passage and creates an appropriate illustration

6.2 Temporal Consistency of I2V model

Here we provide different video sample for passage with car driving as it's main focus.

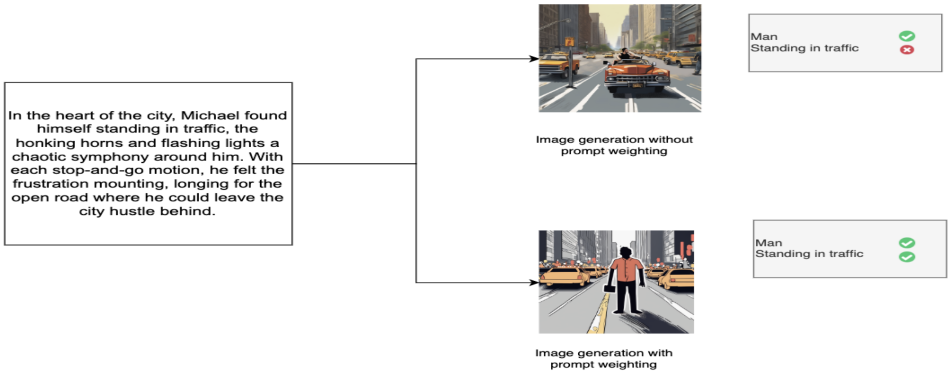
The passage for the video in Figure 13 is "In the dead of night, Sarah found herself behind the wheel of her old Mustang, the road stretching out endlessly before her. With the windows down and the



Along the rugged terrains of northern China, thousands of laborers toiled under the harsh command of the Ming dynasty to extend the Great Wall. The vast landscape was dotted with workers carrying stones and bricks, enduring the biting wind. The wall, a formidable barrier of stone and earth, snaked across mountains and valleys, a testament to human endurance and architectural prowess. Each segment stood as a bulwark against invasions, under the watchful eyes of the empire's sentinels

Image generation without prompt weighting
Great wall of china ✓
Labourers ✗
Image generation with prompt weighting
Great wall of china ✓
Labourers ✓

Figure 10: **Passage-Image alignment for Historical Passages**



In the heart of the city, Michael found himself standing in traffic, the honking horns and flashing lights a chaotic symphony around him. With each stop-and-go motion, he felt the frustration mounting, longing for the open road where he could leave the city hustle behind.

Man ✓
Standing in traffic ✗
Image generation without prompt weighting
Man ✓
Standing in traffic ✓
Image generation with prompt weighting

Figure 11: Prompt Image alignment for Contemporary Fiction

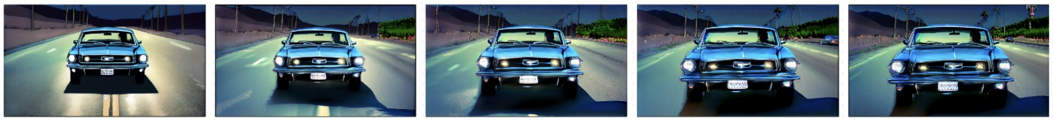


Figure 12: Driving car video from a fictional passage

radio playing softly, she drove aimlessly, seeking solace in the rhythmic hum of the engine and the cool night air rushing past.". As we can infer from the image, we see a Mustang car driving in the dead of the night.The frames are consistent with the expected passage.

7 Conclusion

7.1 Observations from results

- In general, prompt weighting and FreeU parameters are set manually. Using LLM's for this has enhanced the semantic alignment of the passages with the generated image.
- **Guidance scale vs performance**: Guidance scale in diffusion model determines the alignment with the prompt text. If lower values are provided, we get a highly creative image that is out of alignment with the prompt, and higher values generate a more aligned image as per the text, but perform poorly on highly creative passages (Ex: Fantasy). Hence, we need to set a balanced guidance scale for optimal performance. We picked a guidance scale of 12.5.
- **Denoising steps vs performance**: Adding more number of denoising steps improve the quality of the generated image, but increase the image generation time. For an A100 GPU, we used 200 denoising steps for text to image and 100 denoising steps for text to video.

7.2 Future work

- **Negative prompting**: We explored LLM enhanced prompting for passages along with prompt weighting for positive prompts. In a similar fashion, LLM enhanced prompt weighting for negative prompts can be explored. Adding this to the Chain of thought (COT) prompt method can help avoid unnecessary outputs.
- **Clip score for a broad range of classes**: In our current implementation, we have derived the CLIP score for only a limited number of classes from the Kinetics 400 labels. However, for future iterations of our project, we intend to expand this scope significantly. Our plan involves leveraging a broader range of classes from the Kinetics 400 dataset to derive the CLIP score for a more comprehensive set of categories.
- Video scoring: We faced a challenge in evaluating video models with baseline benchmarking methods like FVD scores due the unavailability of the datasets and FVD implementations. We plan to inspect this further and provide an appropriate video benchmark.

7.3 Timeline and Contributions

7.3.1 Timeline

As proposed initially, we started out with model evaluation and initial infrastructure setup. This process took us 3.5 weeks and we were able to run an end to end pipeline from text to video. For the next 3.5 weeks, we explored various methodologies to fine tune the pipeline. Due to lack of availability of training data and resources, we researched on training free approaches that can support this pipeline. We picked prompt weighting and FreeU as the ideal candidates and implemented them. For the last 1 week, we evaluated our approach against the general text to video pipeline from the default passage.

7.3.2 Contributions

- **Abhiram Vadlapatla**: Performed literature survey on existing text to video model pipelines. Specifically deep dived into I2VGenXL model. Setup of initial pipeline of Epicdiffusion for SOL for teammates to contribute. Setup timesformer for video evaluation and performed the accuracy analysis.
- **Manikandan Sundararaman**: Literature survey about the existing text to video and text to image models and their architectures. Explored Chain-of-thought prompting, Prompt weighing and I2VGenXL model. Setup of infrastructure on SOL machine. Worked on EpicDiffusion architecture based on inputs from other teammates and ran it on SOL machine to derive the results. Tried multiple prompts to understand how model behaves. Contributed to drafting of report.

- **Srivalli Kanumuri**: Literature survey about the existing text to video and text to image models and their architectures.Worked on text to image pipeline to generate contemporary images based on kinetics 400 actions. Explored textual inversion to improve the architecture. Worked on evaluation metrics. Contributed to drafting of report. Setup of infrastructure on SOL machine.
- **Dhruv Bindra**: Setup of infrastructure on SOL machine. Literature survey about the existing text to video and text to image models and their architectures. Worked on deriving CLIP score for the generated images. Explored datasets that can be used to fine-tune the model. Wrote prompts based on kinetics 400 labels that can be used to evaluate the model. Worked on evaluation metrics.
- **Kaustubh Kulkarni**: Literature review on existing text to video and text to image models. Explored FreeU model to optimize scaling factors and integrated the same to enhance image generation quality. Explored datasets for model fine-tuning. Contributed to drafting of report. Setup of infrastructure on SOL machine.

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

[2] Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D. and Zhou, J., 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145.

[3] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, pp.24824-24837.

[4] Tong, Z., Song, Y., Wang, J. and Wang, L., 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems, 35, pp.10078-10093.

[5] Si, C., Huang, Z., Jiang, Y. and Liu, Z., 2023. Freeu: Free lunch in diffusion u-net. arXiv preprint arXiv:2309.11497.

[6] Jeong, H., Kwon, G. and Ye, J.C., 2023. Zero-shot generation of coherent storybook from plain text story using diffusion models. arXiv preprint arXiv:2302.03900.

[7] Wallace, B., Gokul, A., Ermon, S. and Naik, N., 2023. End-to-end diffusion latent optimization improves classifier guidance. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7280-7290).

[8] Allingham, J.U., Ren, J., Dusenberry, M.W., Gu, X., Cui, Y., Tran, D., Liu, J.Z. and Lakshminarayanan, B., 2023, July. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In International Conference on Machine Learning (pp. 547-568). PMLR.