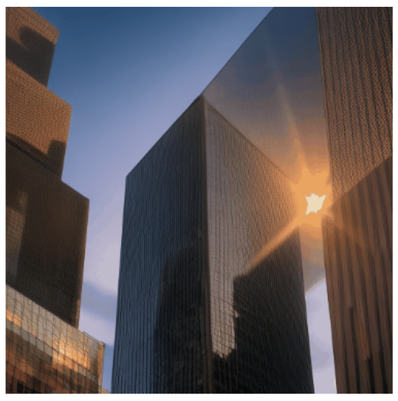
AnimateDiff Lightning: Does faster computation result in lower fidelity?

Team P09

Introduction

Video generation has rapidly evolved with advancements in generative models, yet balancing computational efficiency and output quality remains a challenge. Diffusion-based models like AnimateDiff-Lightning aim to bridge this gap by accelerating video generation while maintaining high fidelity.

Prompt: Skyscrapers reflecting the sunset over a bustling metropolis.



Generated by AnimateDiff (Original Model)

Generated by AnimateDiff-Lightning

Objectives

Research Question: Can AnimateDiff-Lightning achieve comparable or better video quality than the original AnimateDiff while significantly reducing computational costs for the same set of prompts?

Objective:
Through systematic benchmarking and A/B testing across diverse themes and evaluation metrics, we aim to investigate whether the lightning-fast model maintains high fidelity or introduces trade-offs in video quality. Evaluation metrics, inspired by Lin & Yang (2024), include **video-text relevance, appearance distortion, appearance aesthetics, motion naturalness, motion amplitude, and overall quality.**

Outcome Goals:
By analyzing these objectives, we aim to demonstrate that AnimateDiff-Lightning achieves comparable or superior results to the original model, particularly in terms of computational efficiency, while maintaining or improving video fidelity.

Methods

Evaluation Process:

1. Generating videos in GIF format from 100 diverse text prompts covering themes such as sports, nature, and wildlife. Text prompts were carefully crafted to include varying levels of detail, complexity, and thematic diversity to challenge both models equally.
2. Standardizing parameters like the guidance scale, inference steps, and base model style to ensure a fair comparison.

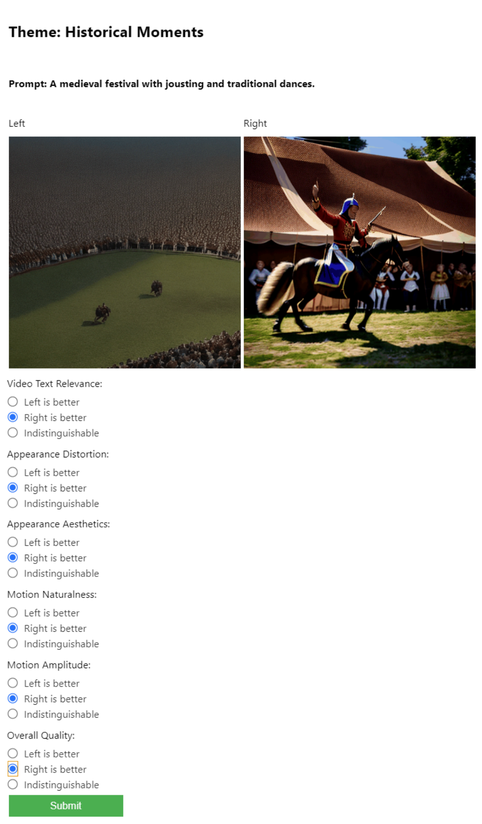
GIFs generated by the two models under different themes.

Evaluation Framework:
To assess the quality of the GIFs, we conducted A/B testing, where evaluators were presented with pairs of GIFs (Model A and Model B) and asked to rate them based on a set of six criteria. The GIFs were presented in a randomized order to ensure evaluators could not identify which model produced each output.

Scoring Criteria:
To move beyond subjective preferences, we used the following six quantitative and qualitative metrics for benchmarking:

1. **Video-Text Relevance:** Measures how well the content of the GIF aligns with the provided text or prompt, ensuring it accurately represents the intended message or concept.
2. **Appearance Distortion:** Assesses the extent of visual artifacts or deformities in the generated animation, focusing on maintaining realistic proportions and minimizing errors.
3. **Appearance Aesthetics:** Evaluates the overall visual appeal of the GIF, considering elements like color harmony, design consistency, and artistic quality.
4. **Motion Naturalness:** Examines the believability and fluidity of the movements, ensuring they mimic realistic or contextually appropriate dynamics.
5. **Motion Amplitude:** Reflects the extent or range of movement in the animation, judging whether the motion is dynamic and contextually fitting without being too subtle or excessive.
6. **Overall Quality:** Provides a holistic evaluation of the GIF, combining all other criteria to judge its coherence, effectiveness, and visual impact.

Results were aggregated and analyzed to identify statistically significant differences between the two models across the six metrics.



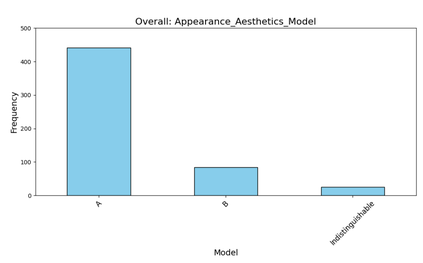
Evaluation

During the evaluation, the following two models were used: AnimateDiff Lightning (Model A) and AnimateDiff (Model B). Participants evaluated GIFs based on the aforementioned six metrics, focusing on both qualitative and quantitative aspects of video quality. Results were aggregated across 100 prompts from various themes such as urban exploration, historical moments, and underwater worlds. After aggregating the results, we came up with a few findings:

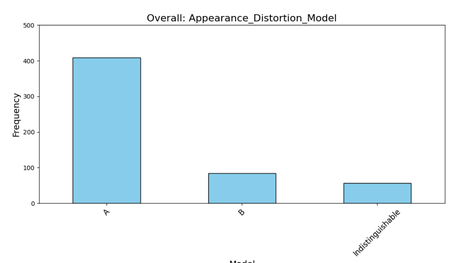
Key Findings:

1. **Overall Preferences:**
   - Model A (AnimateDiff Lightning) was strongly preferred across all metrics, indicating its higher performance overall.
   - However, preferences were less dominant for motion naturalness, motion amplitude, and video-text relevance, where Model A and Model B showed closer results.
2. **Theme-Specific Insights:**
   - Abstract Concepts: Models A and B showed similar preferences for motion amplitude and video-text relevance due to the difficulty in evaluating abstract visualizations.
   - Magical Seasons & Mystical Realms: Indistinguishable results were more common in video-text relevance and motion amplitude.
   - All other themes generally aligned with the overall trends favoring Model A.

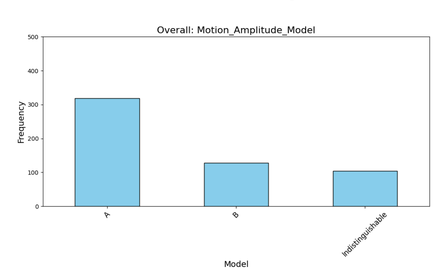
Graphical representation (Overall results):
The following charts highlight the overall results across the six metrics, as well as the overall preferred model after aggregating the six metrics.



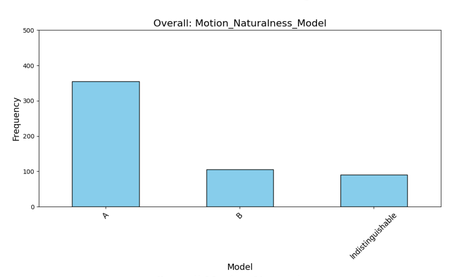
Overall Appearance Aesthetics:
Preference for Model A was significant



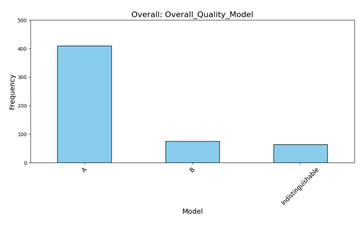
Overall Appearance Distortion:
Preference for Model A was significant



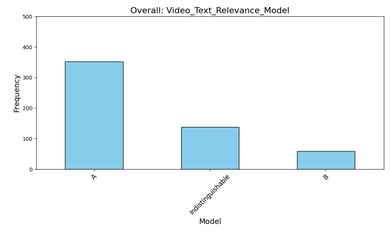
Overall Motion Amplitude:
Model A had a slight advantage,
though preferences were more distributed



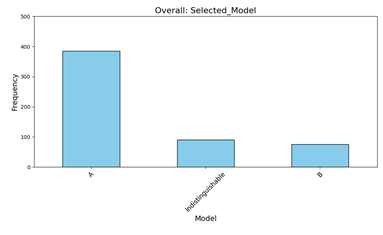
Overall Motion Naturalness:
Model A had a slight advantage,
though preferences were more distributed



Overall Quality:
Preference for Model A was significant



Overall Video-Text Relevance:
Model A had a slight advantage,
though many times GIFs were indistinguishable



Overall Selected Model:
Preference for Model A was significant

These results collectively highlight the efficiency of AnimateDiff Lightning in generating high-quality videos with reduced computational requirements, affirming its potential for broader applications.

Conclusion

Faster computation does not necessarily mean that quality is compromised. In the interesting case of AnimateDiff, the Lightning variant seems to be a much more popular choice amongst respondents in terms of the quality of GIF generated.

The quality of a GIF is measured with six evaluation metrics to truly determine which model generates “better” GIFs, minimizing the influence of personal bias. Additionally, by having multiple participants perform the same evaluation, we believe that any form of bias in a single respondent is suppressed by the others, producing rather reliable results.

Therefore, it is safe to conclude that AnimateDiff Lightning is essentially a strict upgrade to the original model, being faster yet producing higher fidelity GIFs in the process.

Acknowledgements

We would like to credit the relevant papers that we referenced during the course of this project:

1. Lin, S., & Xiao, Y. (2024, March 19). Animatediff-lightning: Cross-model diffusion distillation. arXiv.org. https://arxiv.org/abs/2403.12706
2. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., & Dai, B. (2023, July 10). AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. arXiv.org. https://arxiv.org/abs/2307.04725
3. Zhou, Y., Wang, Q., Cai, Y., & Yang, H. (2024, October 20). Allegro: Open the black box of Commercial-Level Video Generation model. arXiv.org. https://arxiv.org/abs/2410.15458

Yow Siao Kang (20246474), Chan Jie Ru (20246472), Yee Jia Zhen (20246483), Ong Wei Xiang (20246473), Huang Jiayi (20246444)