Character Consistent Diffusion Models for Visual Storylines

Carnegie Mellon University

Srivathsan Badrinarayanan, Arvind Car, Radheesh Sharma Meda

Introduction

This project explores novel methods for generating sequential images that form coherent visual storylines while maintaining consistent character representations. We evaluate three approaches:

1. **Prompt-to-Prompt:** Progressive image generation using prior frames.
2. **Prior-Caption Diffusion:** Iterative caption generation guiding subsequent images.
3. **Prior-Latent Diffusion:** Latent space alignment for temporally consistent image generation.

Our goal: Develop approaches that enhance storyline fidelity while ensuring visual consistency in character representation.

Evaluation

Metrics:

Text-Image Alignment: How well generated images match input prompts.

Character Consistency: Stability of character features across frames.

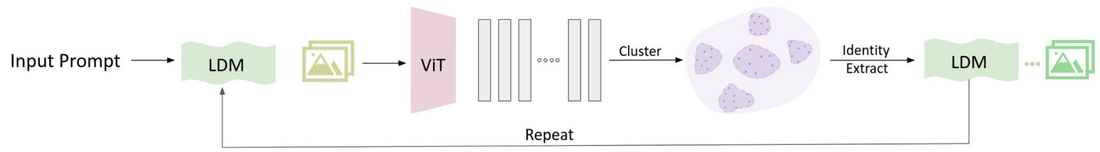
	Human Survey	CLIP Model
Character Consistency	"Rate the main character's consistency across these 5 images"	Directional CLIP similarity (openai/clip-vit-large-patch14)
Text-Image Alignment	"Rate how accurately the image represents the prompt"	CLIP Score (openai/clip-vit-base-patch16)

Evaluation Methods:

CLIP Models: Compute quantitative values (Text-2-Image and Image-2-Image CLIP scores[1]).

Human Surveys: Gather subjective ratings for perceptual quality.

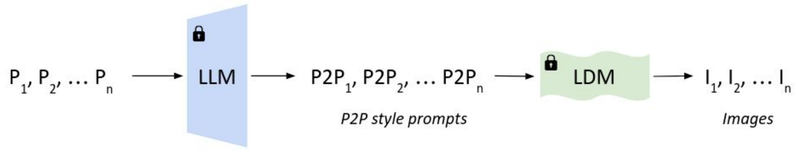
Baseline: The Chosen One[2]



Iterative refinement of generated images based on semantic clustering, guided by cohesive feature extraction and shared characteristics.
Clustering is performed on image embeddings extracted using pre trained DINO v2, guiding the training of the LDM model towards consistent representations.

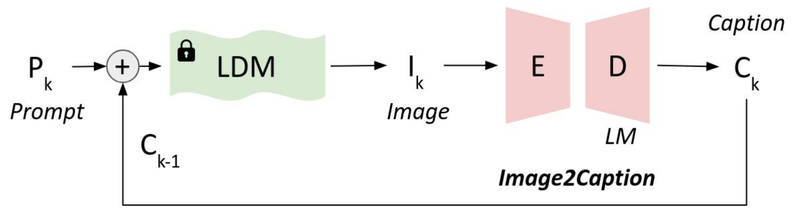
Method 1: Dynamic Base Prompt-to-Prompt

Sequential image generation conditioned on prior frames, with storyline prompts formatted for consistency.



Method 2: Prior-Caption Diffusion

Iterative image generation guided by captions extracted from prior images using a vision captioning[3] + language model pipeline.

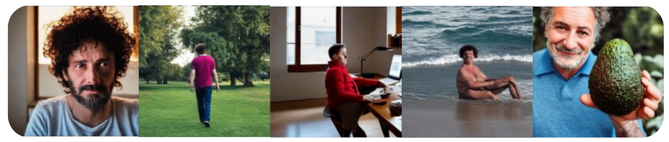


Results

"A photo of a 50 years old man with curly hair…"

in the park | reading a book | at the beach | holding an avocado

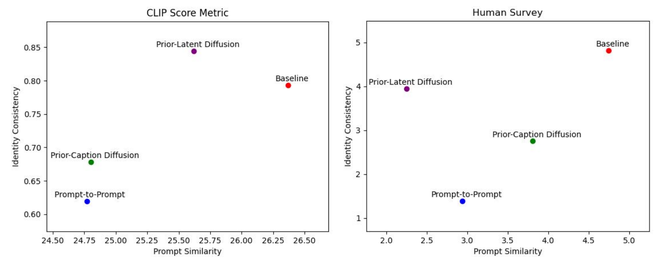
Baseline



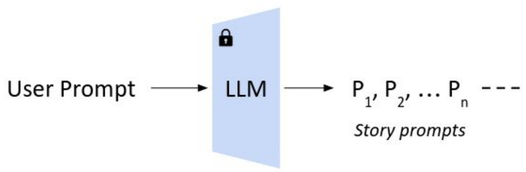
Method 1

Method 2

Method 3

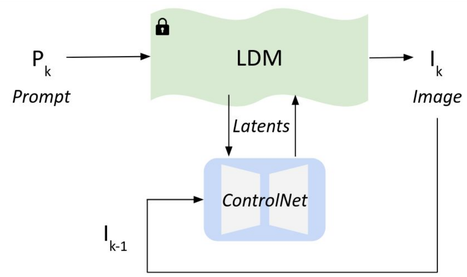


User Input to Story Prompts



Method 3: Prior-Latent Diffusion

Leverages ControlNet[4] to condition subsequent generations directly on previously generated images and the current story prompt.



References

1. Radford, Alec, et al. "Learning transferable visual models from natural language supervision."
2. Avrahami, Y.; Zada, D.; Fried, O.; Aberman, K. The Chosen One: Consistent Characters in Text-to-Image Diffusion Models. IEEE Trans. Pattern Anal. Mach. Intell. 2024, 46, 1234–1246.
3. Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." International conference on machine learning. PMLR, 2022.
4. Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models."