ComicGen: Multimodal AI-powered Tool for Comic Strip

San Jose State University

DATA 298B: MSDA Project 1

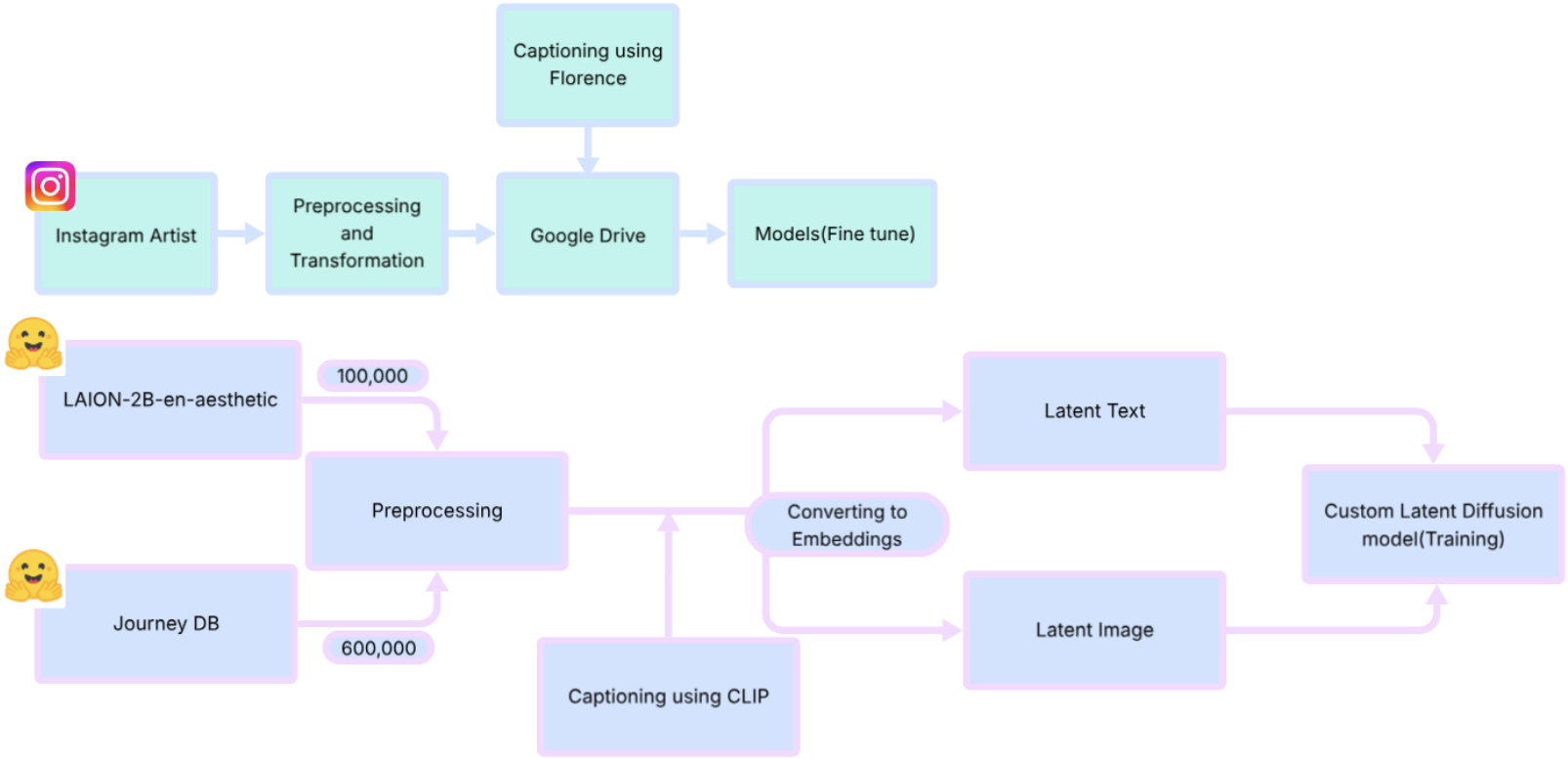
Team 7

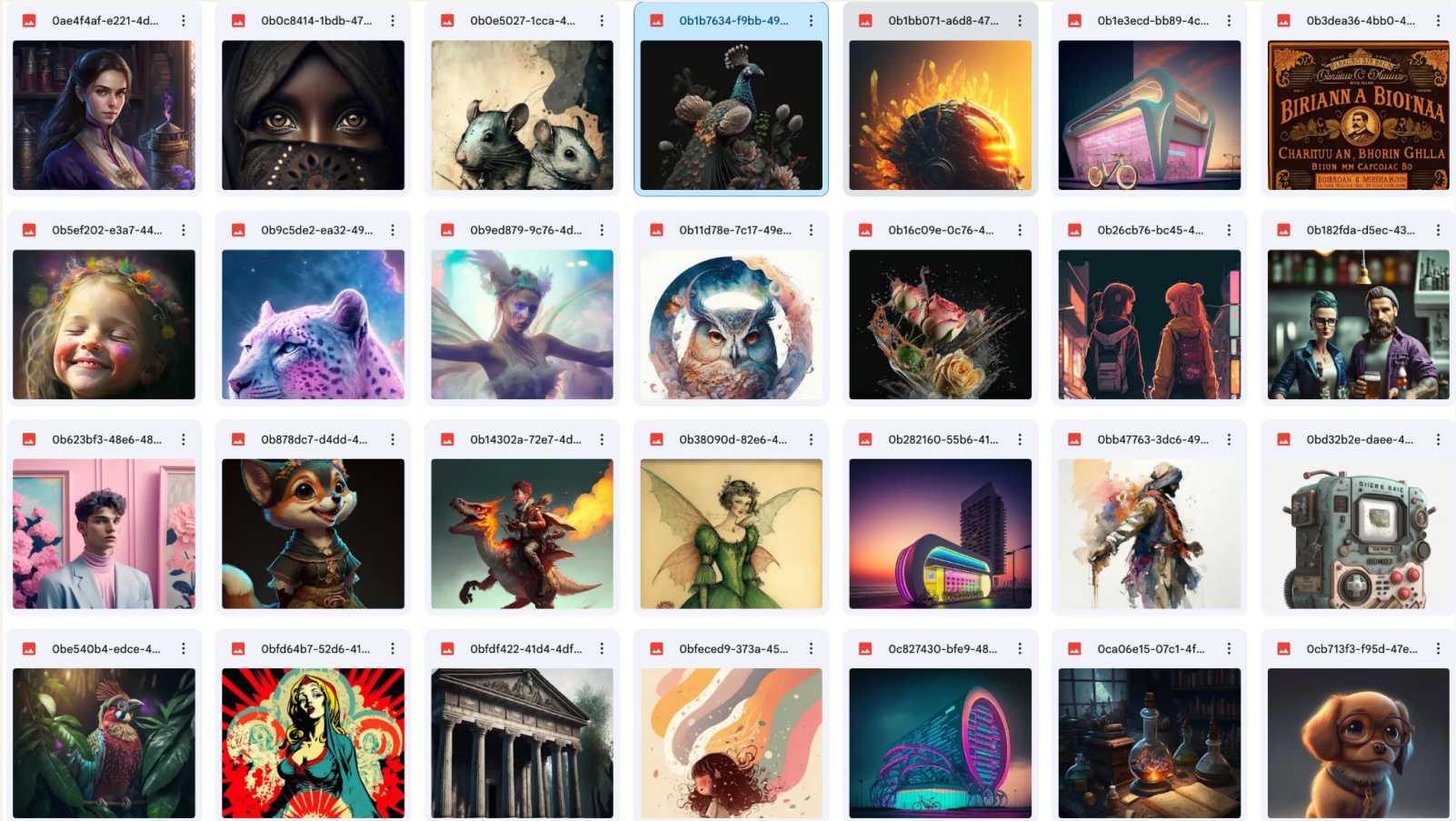
Ananya Varma Mudunuri, Aryama Ray, Nikhil Sarma Gudur, Shreenithi Sivakumar, and Sri Mounika Jammalamadaka

- Stable Diffusion XL (SDXL)
- FLUX
- Lumina
- PixArt-Sigma
- Custom Transformer Latent Diffusion (Spartan model)

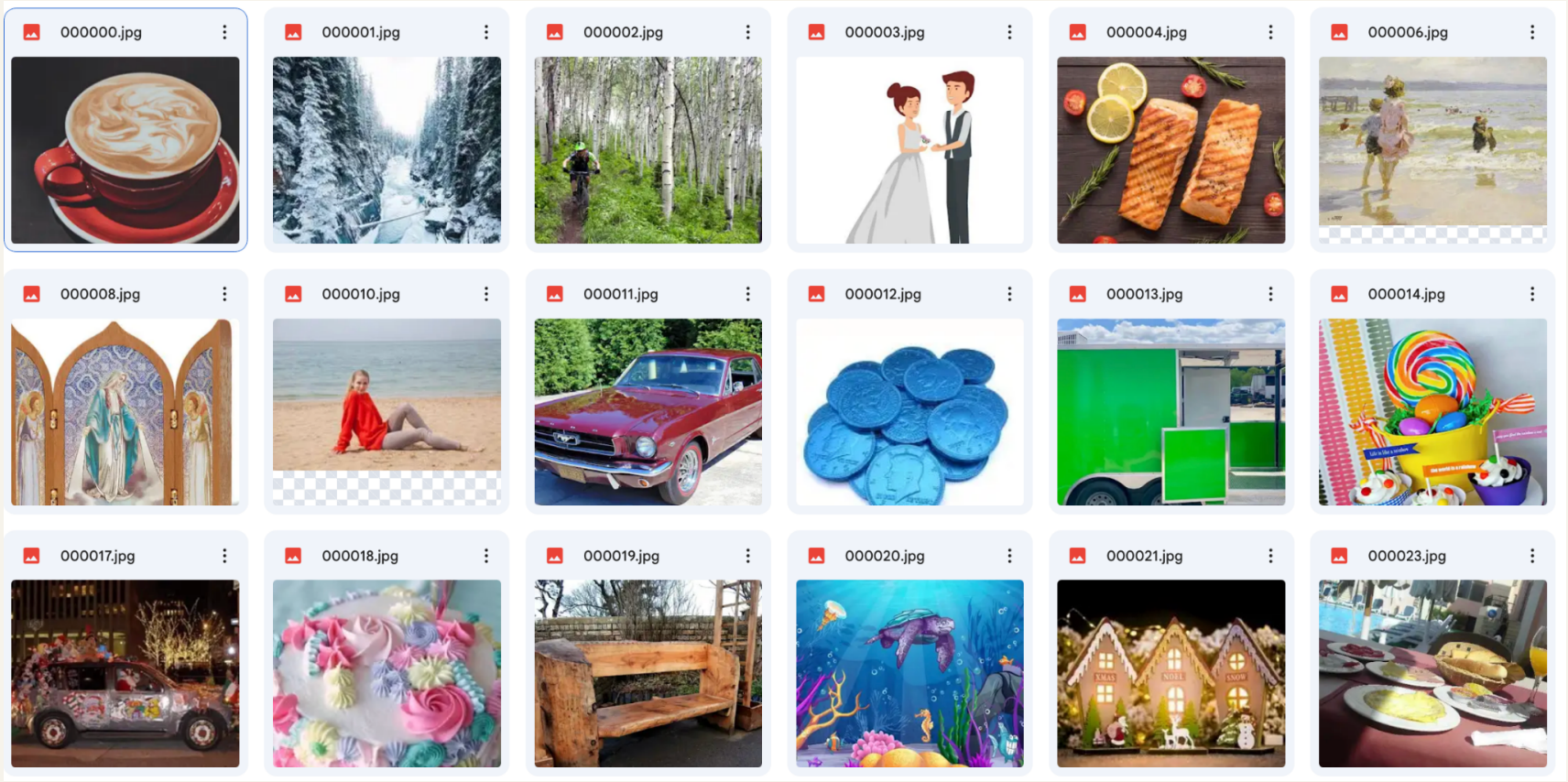
Model Proposals

Data Collection & Preparation





JourneyDB: 600,000 images



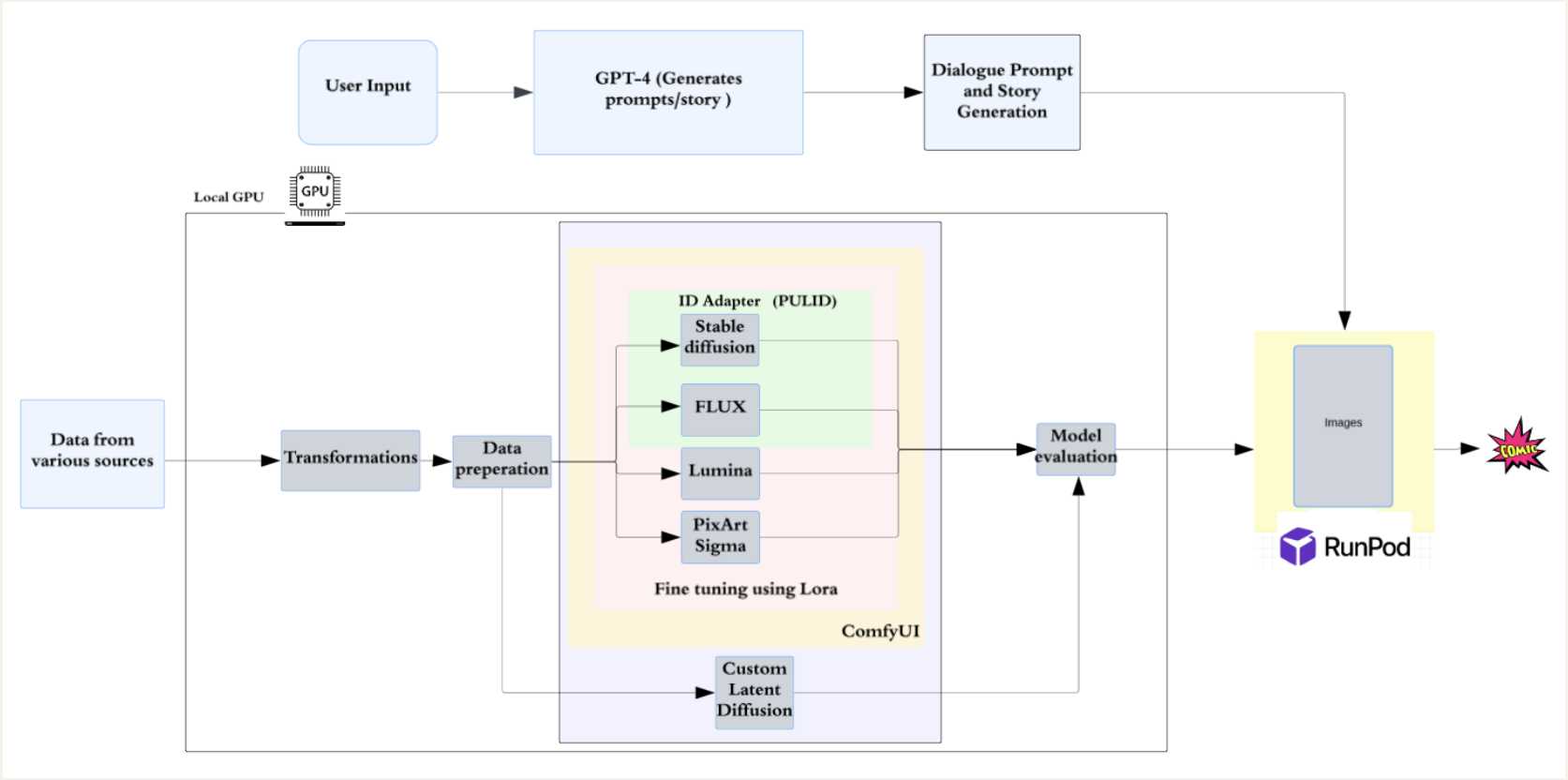
LAION-2B-en-aesthetic: 100,000 images

Instagram images (Fine-Tuning):
30 images

Model Comparison and Justification

Model	Advantages	Justification	Comparison
Stable Diffusion XL	Stable and consistent high-quality image generation. Balanced performance across different text prompts.	Chosen for its reliability in producing visually coherent images with controlled complexity.	Suitable for general text-to-image tasks but lacks intricate detail capture like newer diffusion models.
Lumina	Highly stylized, top-quality image creation Greater control of lighting, texture, and composition High consistency across frames	Employed in the project in order to lend beauty and artful consistency to comic panels Works well with visually rich prompts with mood-specific demands	Compared to SD v1.5 outputs, Lumina has more visually sophisticated results More stylized and expressive but a bit slower than PixArt-Sigma
Flux	Ensures cross-frame consistency in generated images. Reduces scene variations and maintains artistic integrity.	Helps in maintaining a smooth flow across multiple comic panels or animation frames.	Compared to LoRA, Flux targets overall scene coherence rather than specific style adaptations.
PixArt-Sigma	Rapid Image Generation Multi-subject processing and prompt accuracy Outstanding mood, action, and scene execution	Selected for producing full comic panels quickly from story prompts Strikes a balance between visuals and speed for batch or iterative comic-making workflows	Compared to Lumina, PixArt-Sigma is quicker but not quite so stylized Stronger in prompt interpretation compared to base SD models Ideal for iterative outputs
Custom Latent Diffusion	Lightweight (~890M params) and fast training Runs efficiently on a single RTX 490 GPU Transformer-based denoiser yields sharper outputs	Built from scratch for fast and flexible comic generation Allows full control over architecture and training loop Trained on 700k high-quality images (JourneyDB + LAION-2B)	Unlike Stable Diffusion, replaces U-Net with a Transformer-based denoiser More modular and faster than SDXL Optimized for simplicity, custom experimentation, and lower resource usage Ideal for research and iterative comic generation workflows

WorkFlow


User Input
GPT-4 (Generates prompts/story)
Dialogue Prompt and Story Generation
Local GPU
GPU
ID Adapter (PULID)
Stable diffusion
FLUX
Lumina
PixArt Sigma
Fine tuning using Lora
ComfyUI
Custom Latent Diffusion
Data from various sources
Transformations
Data preperation
Model evaluation
Images
RunPod
COMIC

Analysis of Model Execution and Evaluation Results – Before Fine-Tuning

SDXL | Flux | PixArt-Sigma | Lumina

SDXL | Flux | PixArt-Sigma | Lumina

A cheerful barista with long brown hair in a vibrant red apron, pouring coffee in a colorful café

SDXL | Flux | PixArt-Sigma | Lumina

A cheerful barista with long brown hair in a vibrant red apron, serving pastries in a cozy café corner

A cheerful barista with long brown hair in a vibrant red apron, chatting with customers in a lively café

Analysis of Model Execution and Evaluation Results –After Fine-Tuning

SDXL Flux PixArt-Sigma Lumina

A cheerful barista with long brown hair in a vibrant red apron, pouring coffee in a colorful café

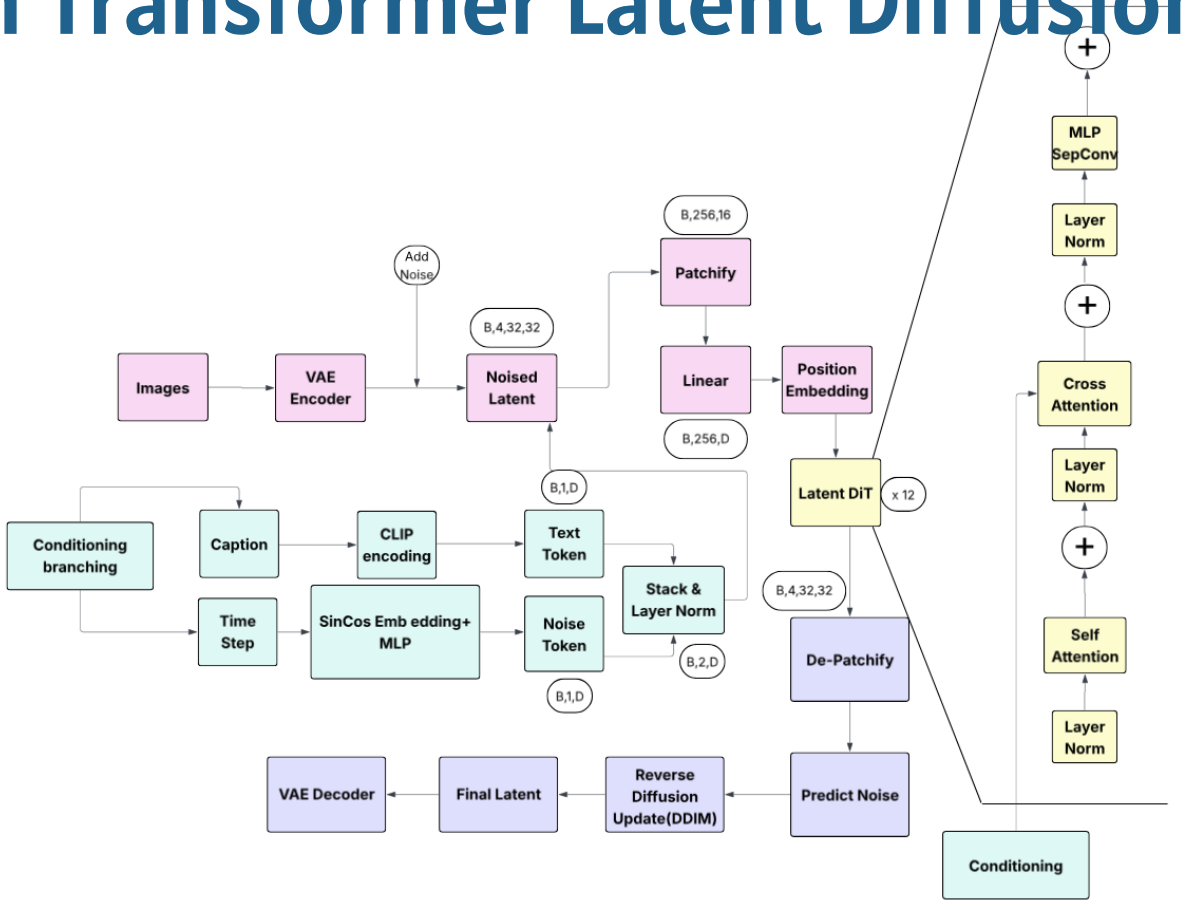
SDXL Flux PixArt-Sigma Lumina

A cheerful barista with long brown hair in a vibrant red apron, serving pastries in a cozy café corner

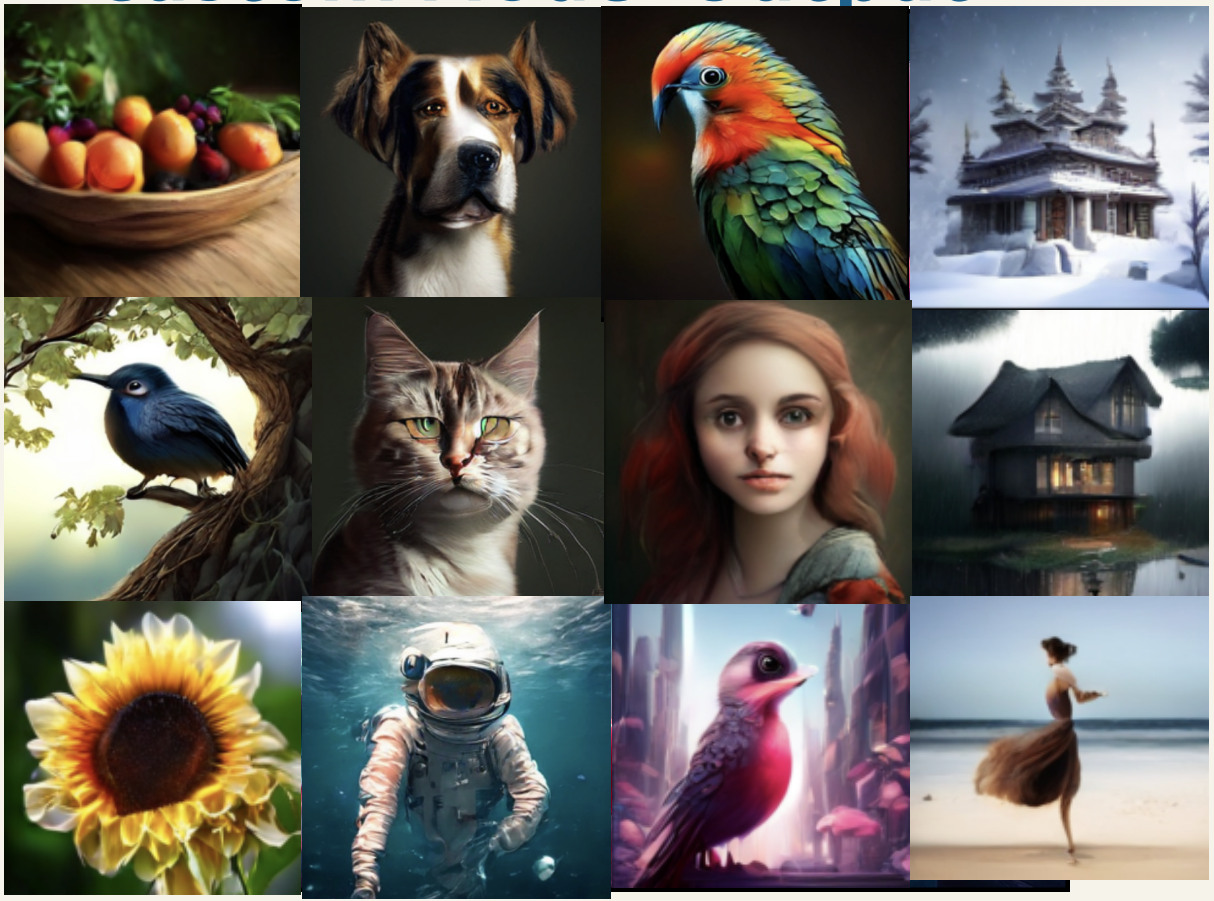
SDXL Flux PixArt-Sigma Lumina

A cheerful barista with long brown hair in a vibrant red apron, chatting with customers in a lively café

Custom Transformer Latent Diffusion Model



Custom Model Output



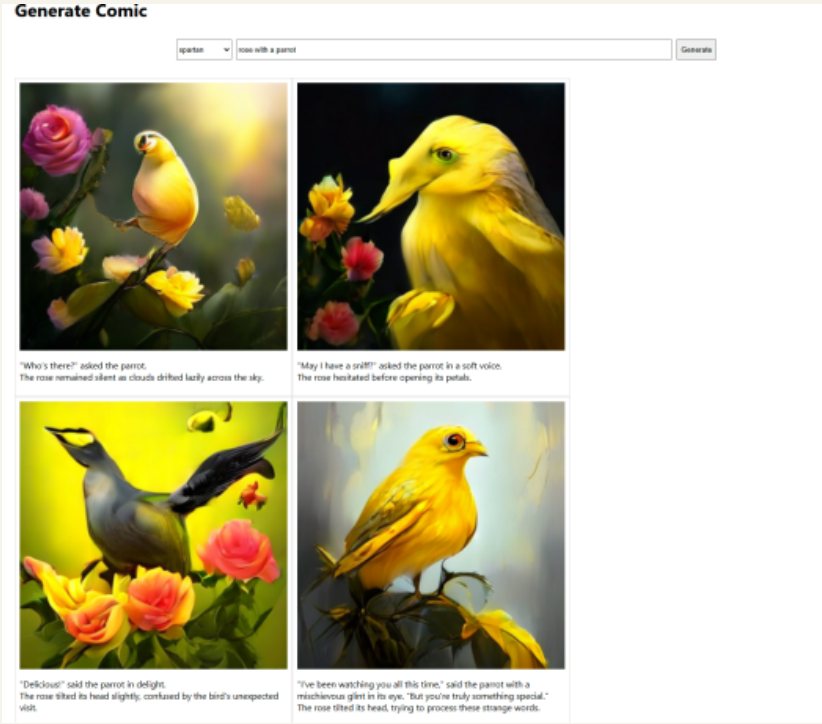
Custom model Output

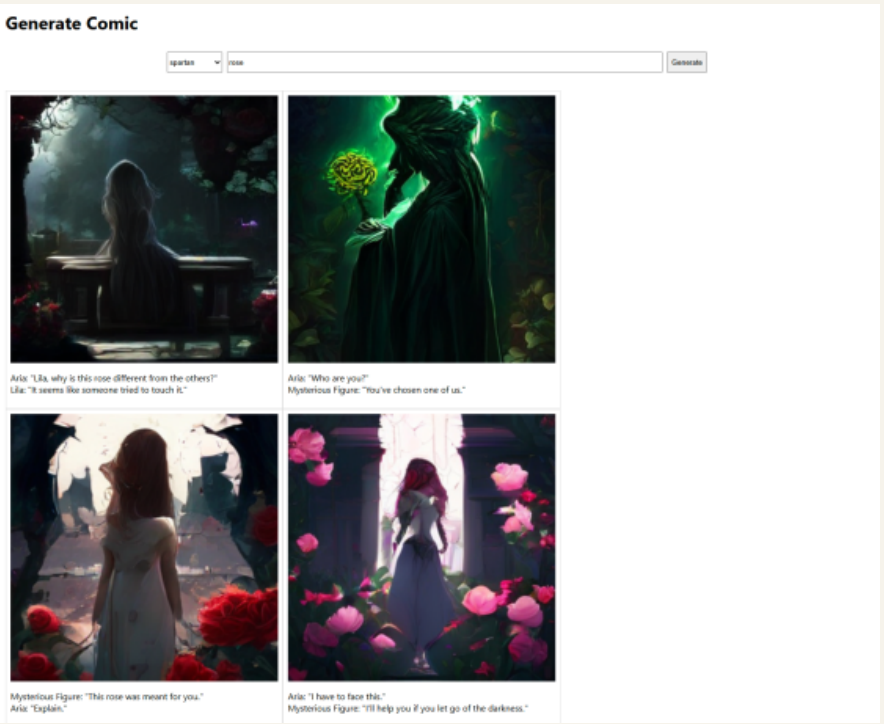
CLIP interpolation:
Cat to Dog

Custom model Output

CLIP interpolation:
Rose to Parrot

Comic Strip generated by custom model





Model Evaluation

CLIP Score

- **Models:** Stable Diffusion XL, Flux, PixArt-Sigma, Lumina, Custom model
- **Description:** Assesses how well generated images semantically match the input text prompts.
- **Interpretation:** Higher CLIP score = Better image-text alignment.

Captioning Accuracy (BLIP)

- **Models:** Stable Diffusion XL, Flux, PixArt-Sigma, Lumina,Custom model
- **Description:** BLIP was used to generate image captions and verify visual-textual coherence during validation.
- **Interpretation:** High-quality captions improve prompt construction and validate alignment between visuals and narrative intent.

Human Evaluation

- **Models:** All (SDXL, Flux, PixArt-Sigma, Lumina,Custom model)
- **Description:** Subjective assessment of story coherence, creativity, visual consistency, and character alignment.
- **Interpretation:** Ensures narrative and stylistic quality across entire comic panels and sequences.

Evaluation Metrics

	CLIP	BLIP
Flux	33.94830922	13.89412185
Base flux	33.71987641	15.21719694
Lumina	34.14467673	14.76545831
Base lumina	31.98541303	14.77437218
Pixart	35.17402212	14.22741761
Base pixart	32.23858774	15.04257967
SDXL	34.82917845	14.30999686
Base sdxl	34.16497111	14.7576645
Custom Model	55.92904762	11.23834021

Summary

Our experiments show that the custom latent diffusion model performs well on fantasy or surreal prompts. However, it struggles with realistic human faces due to limited human-focused training data. Among the fine-tuned models, Stable Diffusion XL (SDXL) emerged as the best performer, producing the most visually coherent and identity-consistent outputs across comic panels, especially after LoRA-based tuning and PuLID integration.

Benefits and Shortcomings

Benefits:

- **SDXL (fine-tuned)** showed excellence in preserving character style and generating detailed backgrounds.
- **Flux model** maintained spatial coherence across multi-panel scenes, supporting narrative consistency.
- **PuLID** improved facial identity retention across poses and emotions.
- **ComfyUI** made training, visual node editing, and model blending intuitive and efficient.

Shortcomings:

- Fine-tuned models (like SDXL) still **struggle with hand anatomy**, requiring more focused data augmentation.
- The **custom model** lacked realism in human face generation due to **limited human data** in training datasets.

Potential System and Model Applications

Education: Use ComicGen to create comics for visual storytelling, lesson materials, and language learning.

Entertainment: Enables indie creators to prototype or fully generate comics without needing drawing skills.

Marketing: Generate comic strips for brand mascots, product stories, and campaigns with visual continuity.

Therapy: Help children or trauma survivors express stories visually.

Accessibility: Create comics using voice or text; support future multilingual generation to preserve cultural narratives.

Experience and Lessons Learned

Data quality and **semantic alignment** were central to model performance.

ComfyUI became a core system for pipeline management and experimentation.

PuLID proved vital to combat **character drift** in long-panel comics.

Training a **custom Transformer-based denoiser** expanded the team's understanding of latent space, noise scheduling, and transformer attention mechanisms.

Tools like **JIRA** and **GitHub** helped streamline teamwork and version control.

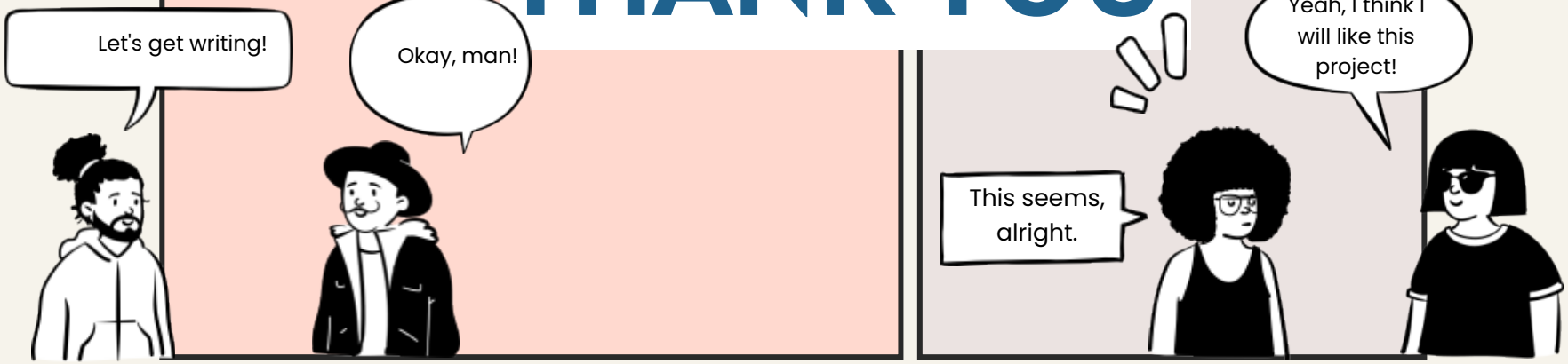
Recommendations for Future Work

- **Multilingual support**: Enable generation in non-English languages using translation layers or multilingual LLMs.
- **Panel Editor**: Add drag-and-drop UI for speech bubbles, scene edits, facial expressions.
- **Mobile deployment**: Use lightweight models (e.g., Distilled SDXL) for mobile or offline use.
- **Character Memory**: Build modules that store personality traits, backstory, and facial features.
- **Motion Comics**: Add capability to auto-generate animated sequences or transitions.
- **User Fine-Tuning**: Let users upload their own styles or faces for personalized mini-models.

Societal Contributions and Impact

- **Cultural**: Helps preserve folk stories and oral traditions in illustrated, shareable form.
- **Educational**: Boosts creativity, digital literacy, and comprehension via visual storytelling.
- **Economic**: Frees freelancers, marketers, and educators from needing full illustration pipelines.
- **Inclusive**: Enables voice-to-comic creation and personal character customization.
- **Creative Equity**: ComicGen gives non-artists and storytellers a tool to express and publish their narratives.

THANK YOU


Let's get writing!
Okay, man!
This seems, alright.
Yeah, I think I will like this project!

Drive Link