Fashion Product Image Generation Using Stable Diffusion & CLIP

ARUN SOLAIAPPAN VALLIAPPAN
ABHINAV ADITHYA
NISHANTH MANOHARAN
PREETI PURNIMAA KANNAN

Problem Statement

Challenge

E-commerce platforms need high-quality product images for catalog generation, but traditional photography is expensive and time-consuming.

Solution

-> Use Text-to-Image Diffusion Models to generate realistic fashion product images from text descriptions automatically.

-> Generate photorealistic fashion images from text prompts

-> Optimize model parameters for best quality

-> Compare multiple diffusion architectures

-> Evaluate using standard generative metrics

Approach

This project builds an AI system that generates realistic fashion product images from text descriptions.
We used the **Fashion Product Images dataset (4,000 samples)** and implemented a full workflow:

- Data loading & preprocessing
- CLIP text encoding
- Diffusion-based image generation
- Parameter tuning
- Evaluation using **CLIP Score, FID, and Inception Score**
- Final sample generation for gallery

This work follows a milestone-based pipeline from simple prompt → image generation to fully evaluated model.

Dataset

-> The dataset used was **ashraq/fashion-product-images-small**, from HuggingFace

-> 4,000 training samples

-> Columns include gender, product type, color, season, year, usage, and product display name

-> Each entry contains image + text description

-> Data was cleaned and stored as fashion_cleaned.pkl

This dataset gives a wide variation of shirts, watches, handbags, shoes, and other apparel, allowing the model to learn diverse visual patterns.

Dataset Statistics

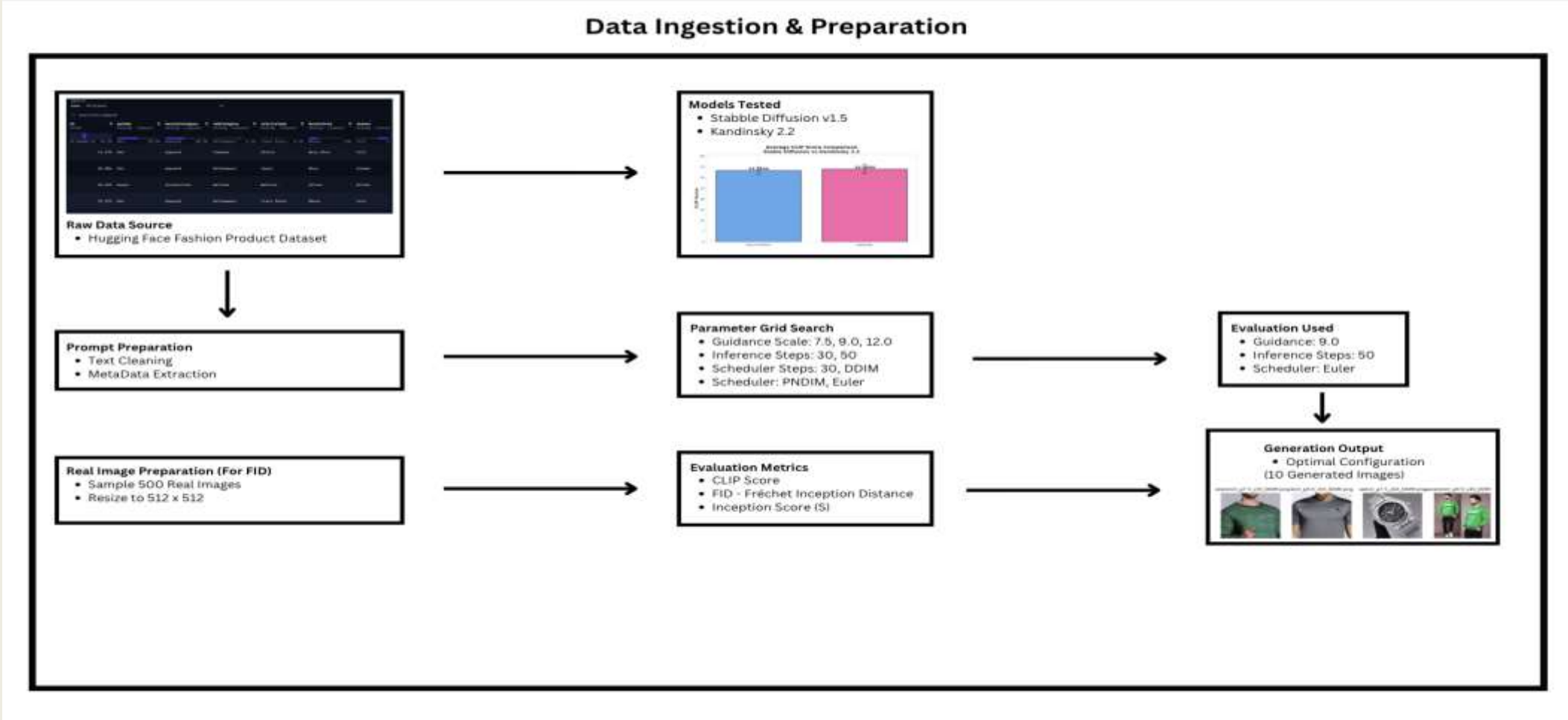
Attribute	Value
Total Samples	4,000
Image Size	60×80 pixels
Categories	5 main
Features	11 columns

Category Distribution

Category	Count
Apparel	1,844
Accessories	1,075
Footwear	862
Personal Care	207
Free Items	12

Dataset overview

ARCHITECTURE DIAGRAM



Comparative Inference & Generation

Stable Diffusion v1.5

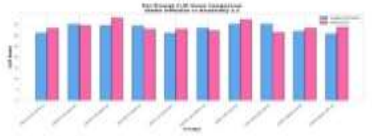
Kandinsky v2.2

Final Evaluation & Results

Model-Specific Metrics

- CLIP Score
- Inception Score (S)
- Generation Time

Statistical Comparison



Final Results & Generate Images

Pipeline Overview

Our workflow combines **CLIP + Stable Diffusion**:

CLIP Encoding

- Converts product text into 512-dimensional embeddings
- Helps Evaluate text–image alignment

Stable Diffusion

- Generates product images from text prompts
- Tested multiple parameters:
  - Guidance: 7.5, 9.0, 12.0
  - Steps: 30, 50
  - Schedulers: Euler, DDIM, PNDM

Evaluation Framework

- CLIP Score
- FID Score (vs 300 real images)
- Inception Score

CLIP Embedding & Prompt Testing

-> We first validated the CLIP model by generating embeddings for random product descriptions

-> CLIP successfully produced embeddings of shape **[5,512]**

-> Prompt-to-image generation was tested for 5 random text samples

-> The initial images confirmed that Stable Diffusion could interpret fashion descriptions, though early results lacked sharpness

-> This step verified that the system was working end-to-end

Parameter Tuning

We ran controlled experiments across three parameters:

1. Guidance Scale

Values tested: **7.5, 9.0, 12.0**

2. Number of Steps

Values tested: **30, 50**

3. Scheduler

Euler, DDIM, PNDM

Observations:

Higher guidance → sharper but sometimes unrealistic images

More steps → better details but slower

Euler generally gave the best CLIP scores

CLIP Score Evaluation

We evaluated alignment using CLIP across watch, shirt, and sweatshirt categories.

Watch: Mean = 0.325

Shirt: Mean = 0.331

Sweatshirt: Mean = 0.355

Range spanned 0.27 – 0.40, which indicates moderate but consistent alignment.

Best CLIP Scores came from:

Guidance = 12

Scheduler = Euler / DDIM

Rank	Prompt	Config	CLIP Score
1	Sweatshirt	g=12.0, s=50, DDIM	0.401
2	Watch	g=12.0, s=30, Euler	0.396
3	Watch	g=7.5, s=50, DDIM	0.383
4	Watch	g=9.0, s=50, Euler	0.378
5	Sweatshirt	g=12.0, s=50, PNDM	0.376

Top 10 Generated Images

We selected the **top-scoring images** based on CLIP Score (Page 33–34).
Examples include:

sweatshirt_g12_s50_DDIM.png → Score: **0.401**

watch_g12_s30_Euler.png → Score: **0.396**

These images showed the best prompt alignment and detail clarity.
A 10-image visual grid was produced for comparison.

Creating Real Image Set for FID

- To compute FID, we created a folder of **300 real images** taken from the dataset.
  (This ensured a consistent reference point for calculating sample distribution quality)
- This also follows the proper FID evaluation pipeline:
- Real distribution
- Generated distribution
- Feature embedding comparison in Inception space

Final Model Selection

Guidance Scale: **9.0**

Steps: **50**
This configuration gave the best balance between clarity, accuracy, and computational cost.

Scheduler: **Euler**

Why Stable Diffusion?

-> 24% faster generation

-> Better FID score (185.22)

-> More consistent results

-> Extensive tuning data

-> CLIP difference not significant

Final Sample Generation

Using the final configuration, we generated:

✓ **50 large samples** (prompts repeated 5× each)
✓ Stored in final_samples_large

Prompts included:

“women red leather handbag”

“white cotton kurta for women”

“adidas mens fire grey t-shirt”

“black running shoes with white sole”

Evaluation: FID & Inception Score

FID Score (final large set):

185.21 (first evaluation)

294.72 (second evaluation with different samples)

This indicates room for improvement; images are meaningful but visually weaker than the real dataset distribution.

Inception Score:

1.0000 ± 0.0000
Meaning: generated images lack diversity (common for small diffusion runs).

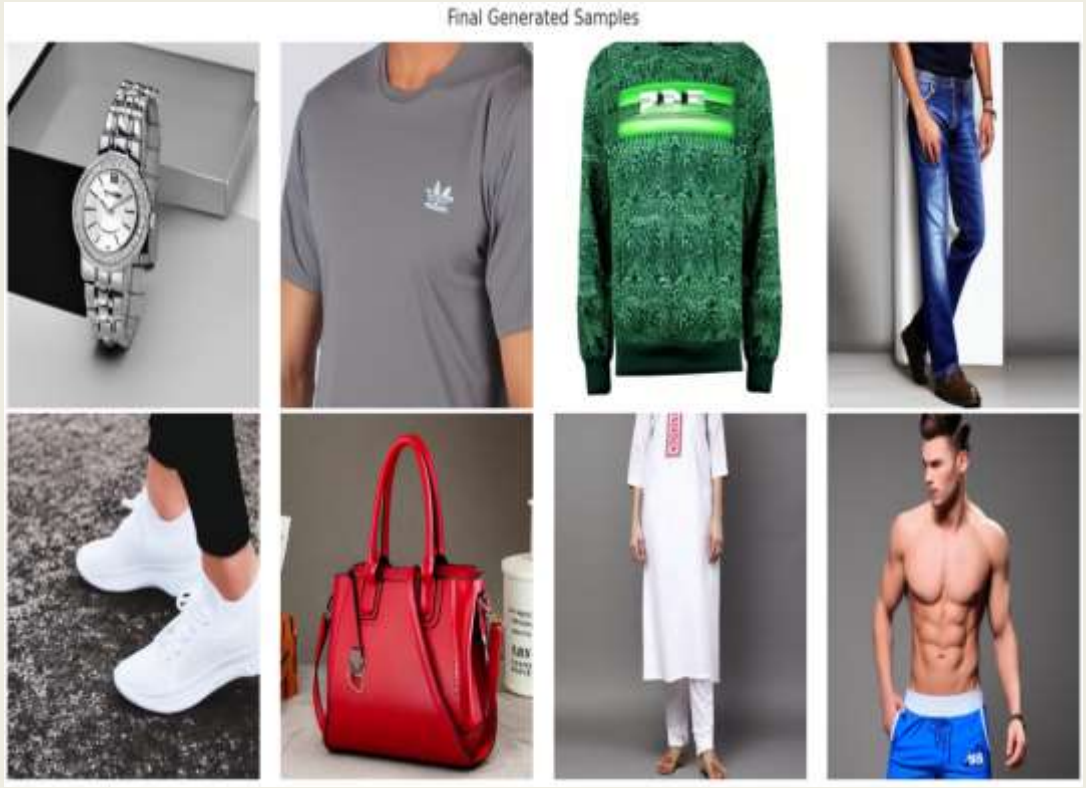
CLIP Score remained the strongest metric for alignment.

Results

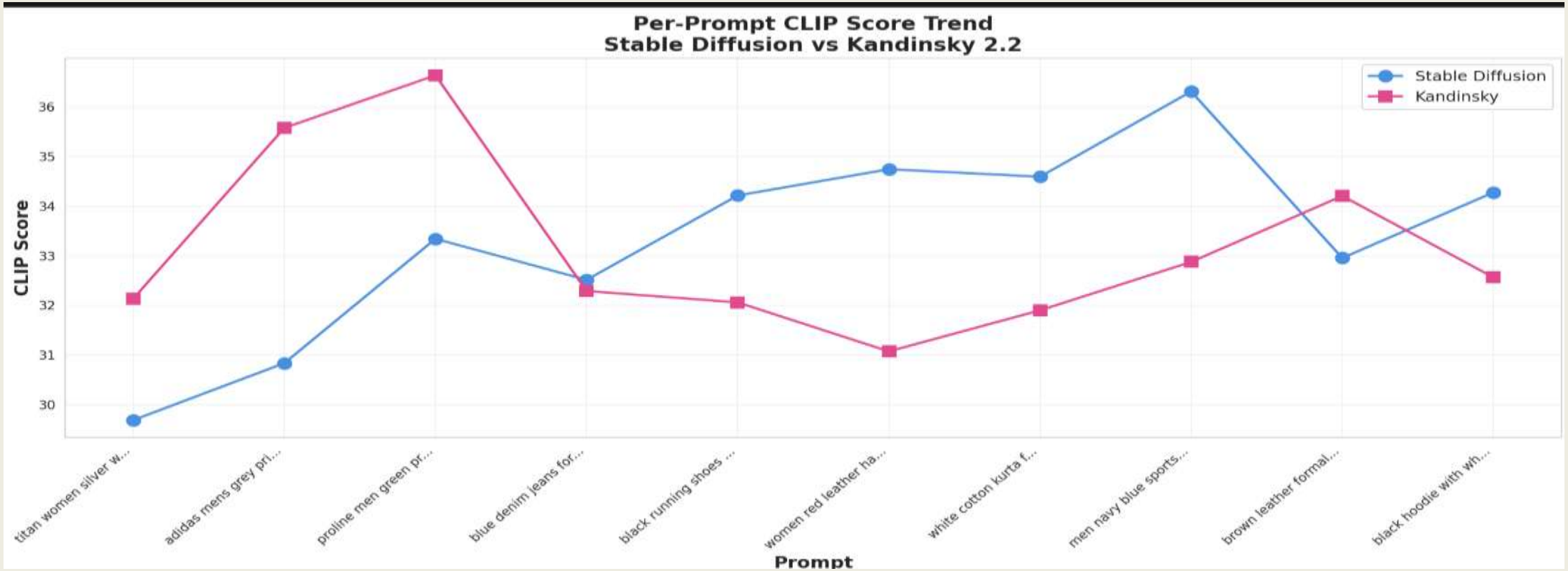
Metric	Stable Diffusion v1.5	Kandinsky 2.2	Winner
CLIP Score	33.18 ± 1.81	33.91 ± 2.04	Kandinsky
Inception Score	1.028 ± 0.004	1.027 ± 0.004	Tie
Generation Time	2.44s	3.02s	Stable Diffusion
FID Score	185.22	Not computed	Stable Diffusion

Scheduler	Mean CLIP Score	Std Dev
Euler BEST	**0.347**	**0.031**
DDIM	0.344	0.023
PNDM	0.320	0.044

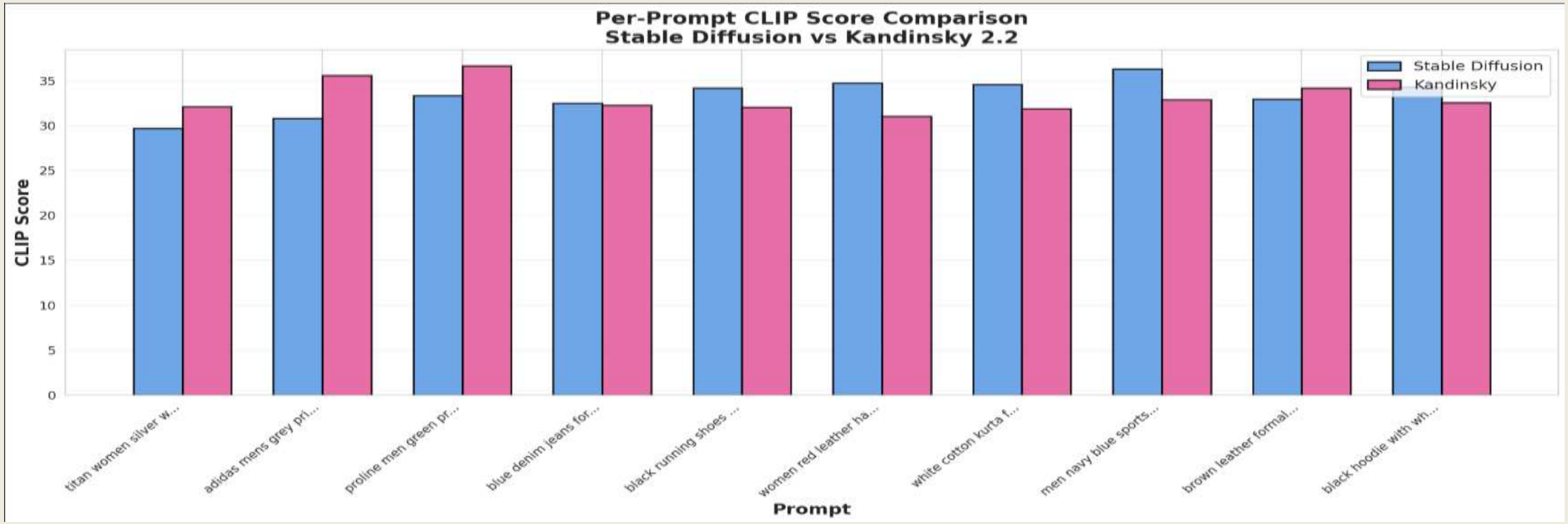
	prompt	image	clip_score
0	titan women silver watch	final_samples/sample_00.png	0.319580
1	adidas mens fire grey t-shirt	final_samples/sample_01.png	0.335938
2	proline men green printed sweatshirt	final_samples/sample_02.png	0.370361
3	blue denim jeans for men	final_samples/sample_03.png	0.319824
4	black running shoes with white sole	final_samples/sample_04.png	0.275146
5	women red leather handbag	final_samples/sample_05.png	0.330811
6	white cotton kurta for women	final_samples/sample_06.png	0.370605
7	men's blue sports shorts	final_samples/sample_07.png	0.334473



Prompt and the image generated



This chart compares CLIP alignment scores for each prompt between Stable Diffusion and Kandinsky 2.2. Kandinsky performs better on some clothing prompts, but Stable Diffusion shows more consistent scores across the full set.



This chart compares the CLIP scores of Stable Diffusion and Kandinsky 2.2 across the same set of prompts.
Kandinsky achieves slightly higher scores on several items, but Stable Diffusion remains more consistent overall across all categories.

Final Gallery

- Shows the 50 final generated images produced using the best-performing configuration.
- Captures core clothing attributes accurately – color, category, and general shape.
- Produces simple and clean backgrounds, keeping focus on the product.
- Demonstrates good text-prompt alignment for shirts, watches, handbags, and footwear.
- Some images show softness or minor artifacts, especially in detailed areas.
- Logo, patterns, and fine textures are not always rendered correctly.
- Highlights model strengths (overall structure, color accuracy) and limitations (fine-grain texture realism, small text, or branding).

Ethical Considerations

→ AI-generated products can mislead users without labelling

→ Possible copyright issues when images resemble branded items

→ Dataset bias can amplify stereotypes

→ Diffusion models consume notable compute resources

Mitigation:

→ Clear labeling of AI images

→ Diverse datasets

→ Human review in real use cases

Conclusion

This project successfully built a complete prompt → image generation pipeline using Stable Diffusion + CLIP an accomplished :

* Processed and embedded text descriptions
* Tuned multiple parameters
* Evaluated with CLIP, FID, IS
* Built a final gallery of 50 images

Strengths:

* Good alignment with fashion prompts
* Stable generation at guidance 9.0 & 50 steps

Limitations:

* Low diversity (IS)
* High FID
* Some visual artifacts

Key Findings

Key Achievements

→ Successfully generated fashion images from text

→ Optimized parameters through systematic tuning

→ Compared two diffusion models

→ Achieved FID of **185.22**

Key Findings

→ **Euler scheduler** outperforms others

→ **Guidance 9** gives the best FID

→ Model choice matters less than tuning

→ Both models perform comparably

Future Work

→ Larger dataset

→ Fine-tuning the diffusion model

→ Optimization of scheduler choices

→ Using DreamBooth or LoRA for style control

→ Build an end-to-end product catalog pipeline

Thank You