LIKE: A Comparative Study of AI Image Generators Through CLIP Semantics and Multi-Faceted Output Analysis

Rohit George

Intro and Motivation

Problem Context

- Rapid growth of AI image generators (e.g., DALL·E, Midjourney, Stable Diffusion) has created a need for better comparative understanding.
- Despite similar prompts, models often produce visually and semantically different outputs.
- Users lack tools to systematically evaluate **how** and **why** models differ in their output.

Motivation

- Evaluate whether different models interpret prompts similarly or diverge in style, consistency, and detail.
- Understand what makes one model "more creative" and another "more aligned" with the text prompt.
- Go beyond subjective visual inspection to **quantify** model behavior using embedding-based techniques.

Key Questions

- Do models interpret prompts in semantically similar ways?
- Which models are more creative, consistent, or faithful to prompts?
- Can image embeddings reveal meaningful trends in generation behavior?

Text-to-Image Models

- **DALL-E 3 (OpenAI):**
  - Uses a CLIP-based text encoder and a diffusion decoder to generate images from text. Known for strong semantic alignment and photorealistic results.
- **Midjourney:**
  - A proprietary model focused on artistic and stylized imagery. While less transparent, it emphasizes visual aesthetics and often departs from strict realism.
- **Stable Diffusion (Stability AI):**
  - A latent diffusion model that generates images by denoising in a compressed latent space, offering high flexibility and open-source accessibility.
- **Kandinsky (by Sber AI):**
  - A two-stage diffusion system that first generates a semantic image layout, then refines it into a final image. Combines diffusion and transformer methods.
- **Adobe Firefly:**
  - Trained on licensed and public domain content, Firefly prioritizes prompt safety and commercial use. Leverages diffusion and Adobe’s creative cloud ecosystem.

Tools and Technologies

- **CLIP (Contrastive Language–Image Pretraining):**
  - Trained on hundreds of millions of image–text pairs, CLIP learns to embed both **text and images into a shared latent space**. It does this by aligning textual and visual representations such that semantically similar content (e.g., a caption and its image) are close in that space.
  - This enables meaningful **image-text comparisons**, clustering, and semantic analysis across models.

- **t-SNE (t-distributed Stochastic Neighbor Embedding):**
  - A **nonlinear dimensionality reduction** technique used to visualize high-dimensional data (like CLIP embeddings) in 2D. It preserves **local similarity**, meaning nearby points remain close in the 2D plot, making it ideal for identifying small clusters or model-specific styles.
- **PCA (Principal Component Analysis):**
  - A **linear dimensionality reduction** method that projects data into directions of greatest variance. Useful for global structure and identifying the **overall spread** of model outputs in CLIP space.

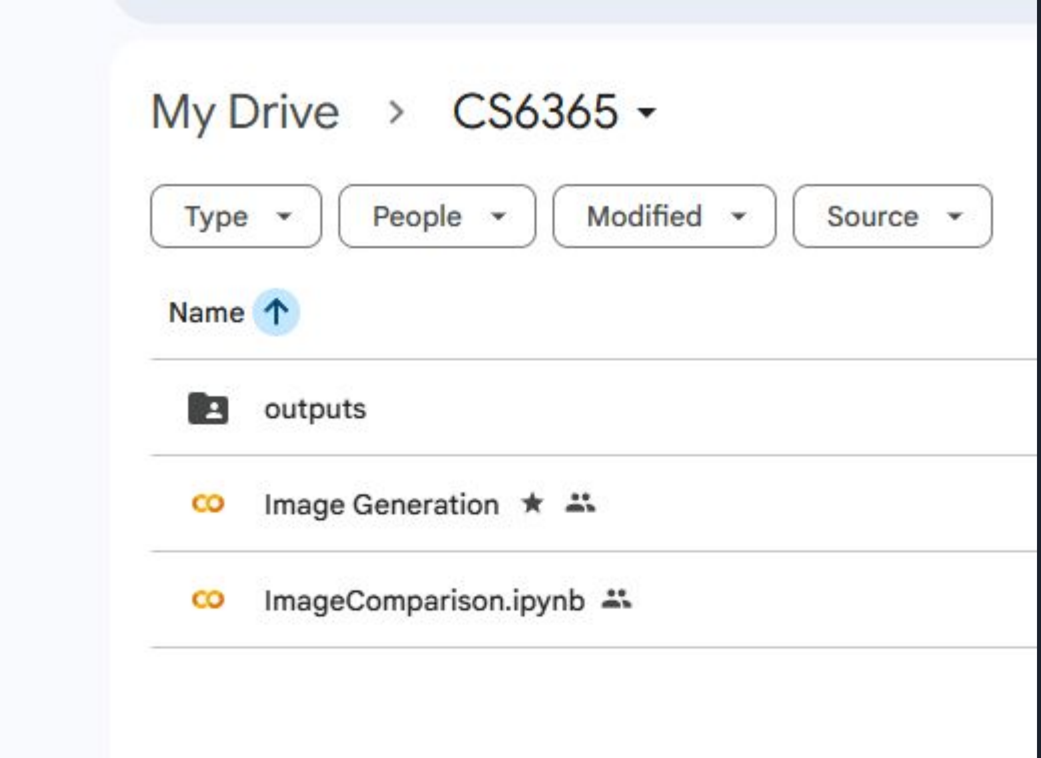
- **UMAP (Uniform Manifold Approximation and Projection):**
  - A **nonlinear projection** method like t-SNE but preserves more of the **global structure**. It is ideal for examining relationships between clusters and discovering broader trends across prompts and models.

Dataset

- Generated 5 different prompt categories
- Generated 4 prompts per category, for a total of 20 prompts.
- Ran these prompts through each image three times, for a total of 300 images.
  - Stable Diffusion and Kandinsky through Hugging Face.
  - DALL-E, Midjourney, and Adobe Firefly through web interface.

Google Drive & Colab

- Hosted on Google Drive
    - Used Colab
- Uploaded to GitHub



My Drive › CS6365

outputs

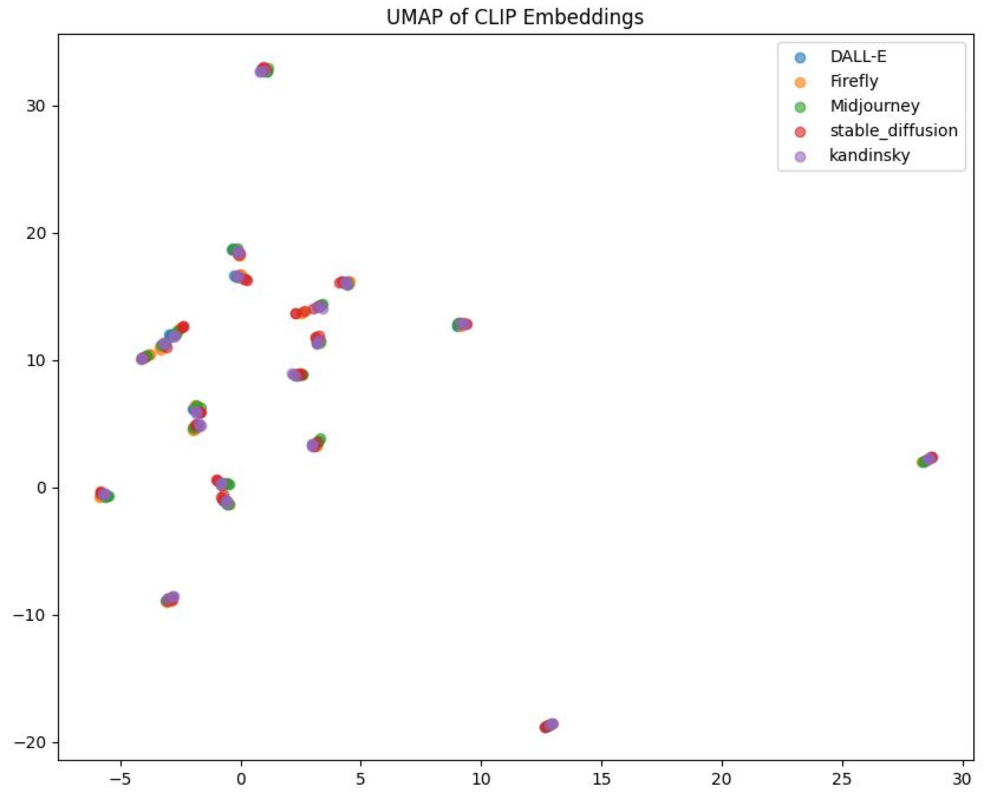
Image Generation

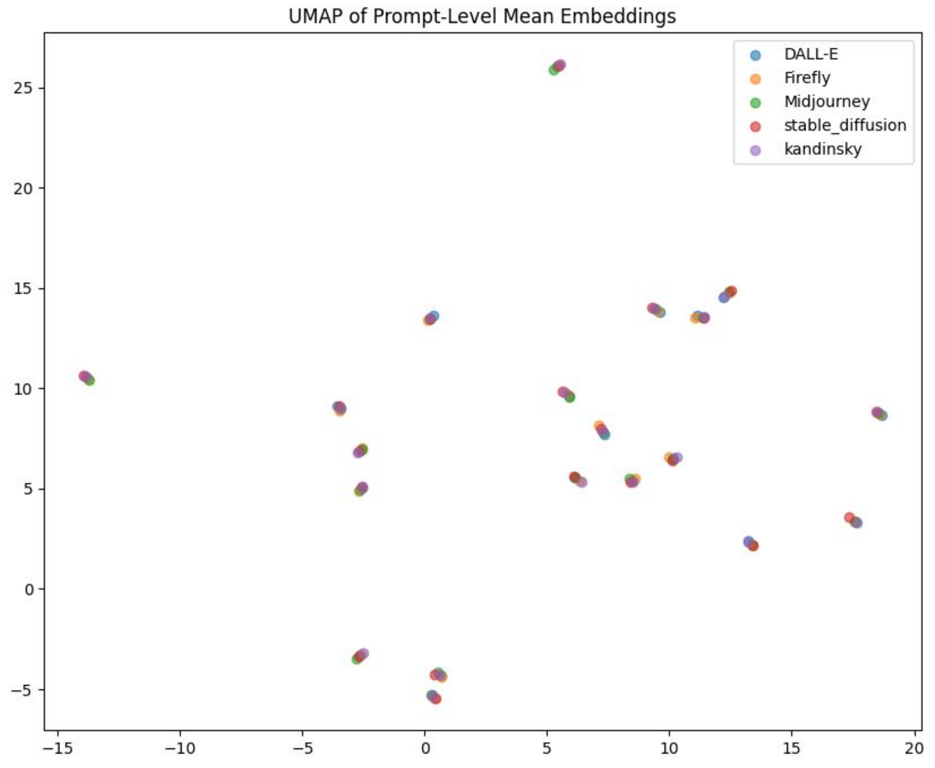
ImageComparison.ipynb

Evaluation Overview

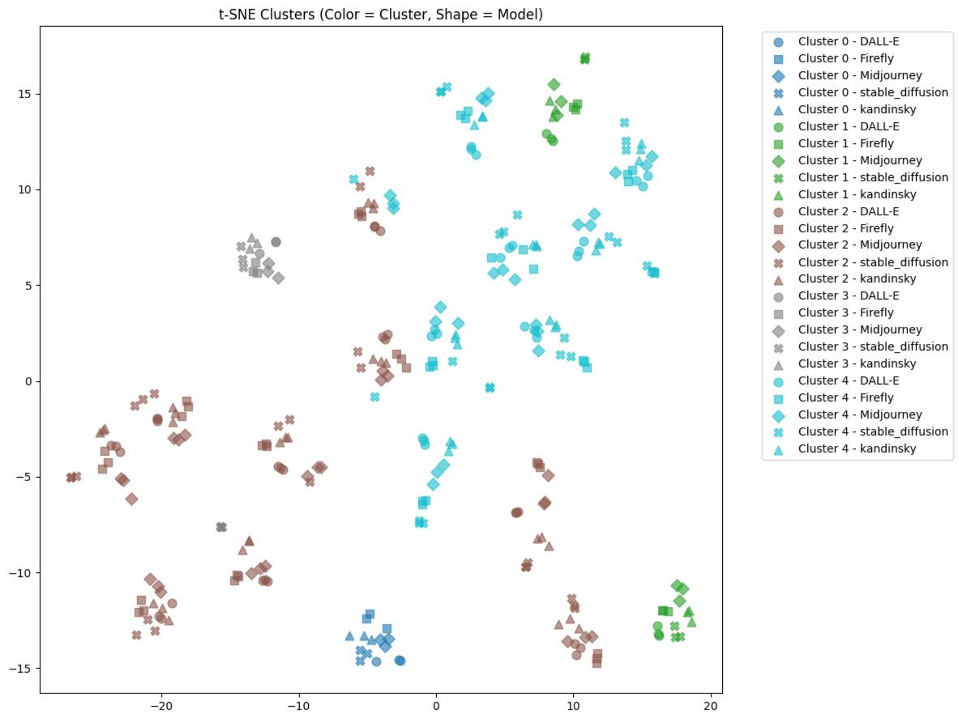
- Compared models in three different cases::
  - Global Comparison (All prompts across all models)
    - Allowed me to observe the overall distribution of image embeddings across models in semantic space, identify whether certain models exhibit broader stylistic variance, and assess clustering behavior related to prompt content or model tendencies.
  - Inter-Model Comparison (Same prompt across different models)
    - Allowed me to evaluate how similarly different models interpret the same prompt, using mean embedding similarity and t-SNE/UMAP visualizations to detect consistent or divergent semantic outputs across models.
  - Intra-Model Comparison (Same prompt, same model, multiple trials)
    - Allowed me to measure the consistency and creativity of each model by analyzing how varied their outputs were across multiple generations of the same prompt, using intra-prompt CLIP similarity and variance metrics.

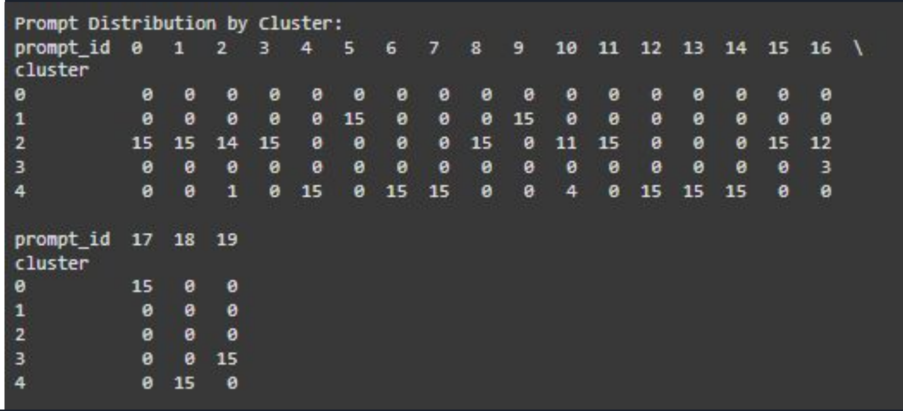
Global Comparison





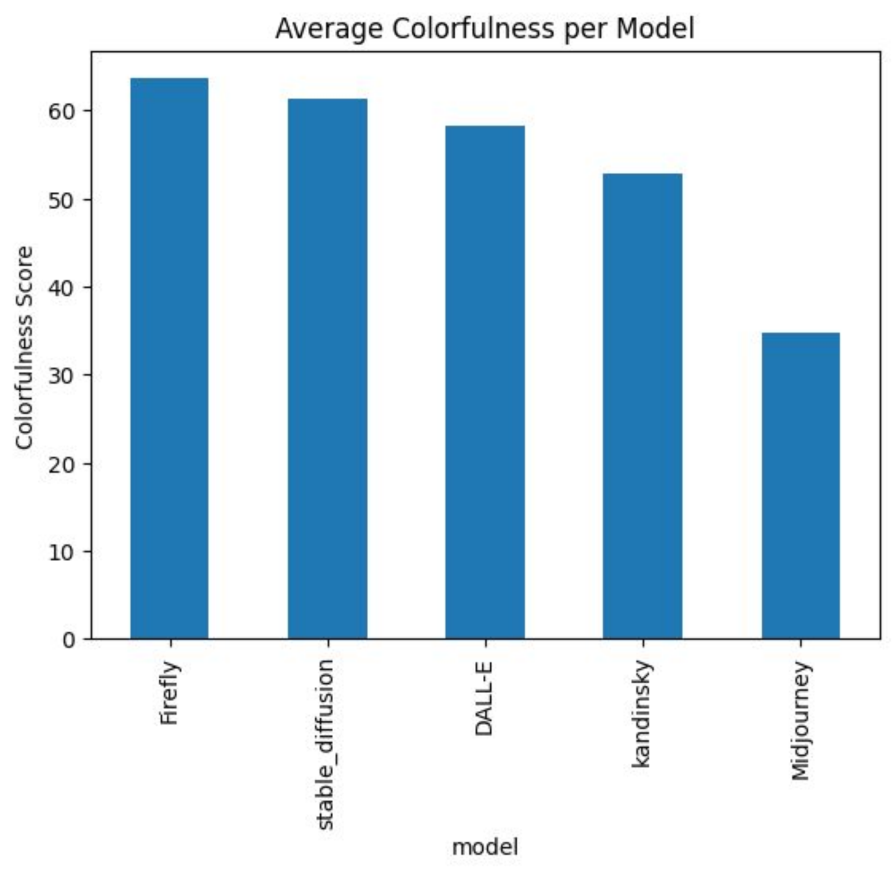
Global Comparison



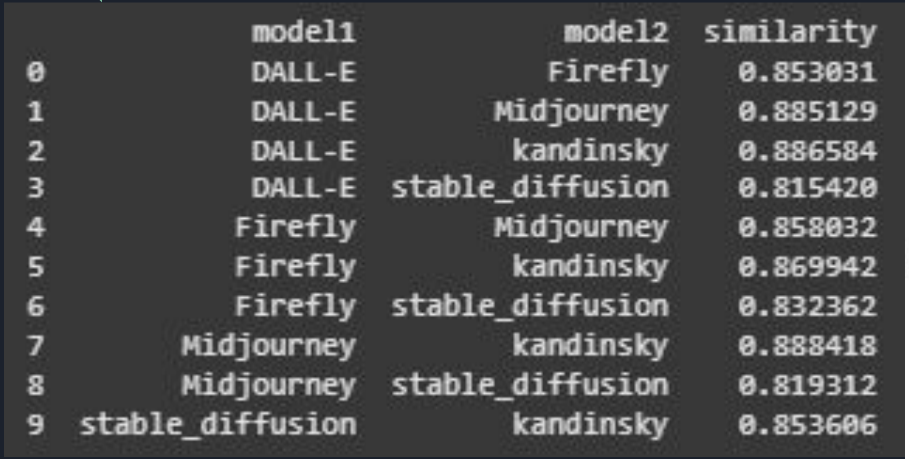


```
Prompt Distribution by Cluster:
prompt_id  0   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  \
cluster
0           0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
1           0   0   0   0   0  15   0   0   0  15   0   0   0   0   0   0   0
2          15  15  14  15   0   0   0   0  15   0  11  15   0   0   0  15  12
3           0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   3
4           0   0   1   0  15   0  15  15   0   0   4   0  15  15  15   0   0

prompt_id  17  18  19
cluster
0          15   0   0
1           0   0   0
2           0   0   0
3           0   0  15
4           0  15   0
```

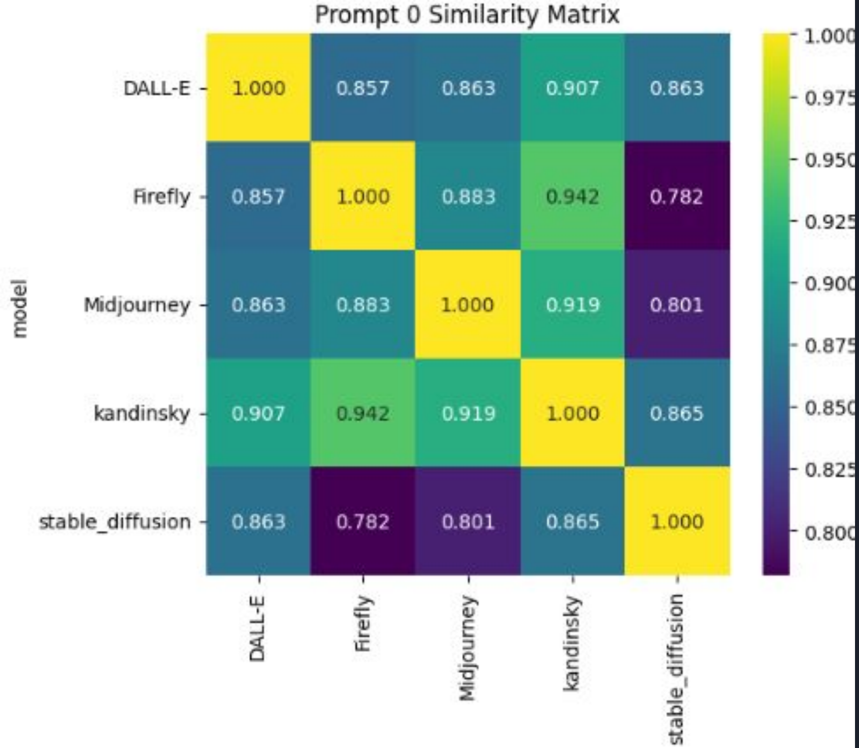
Global Comparison



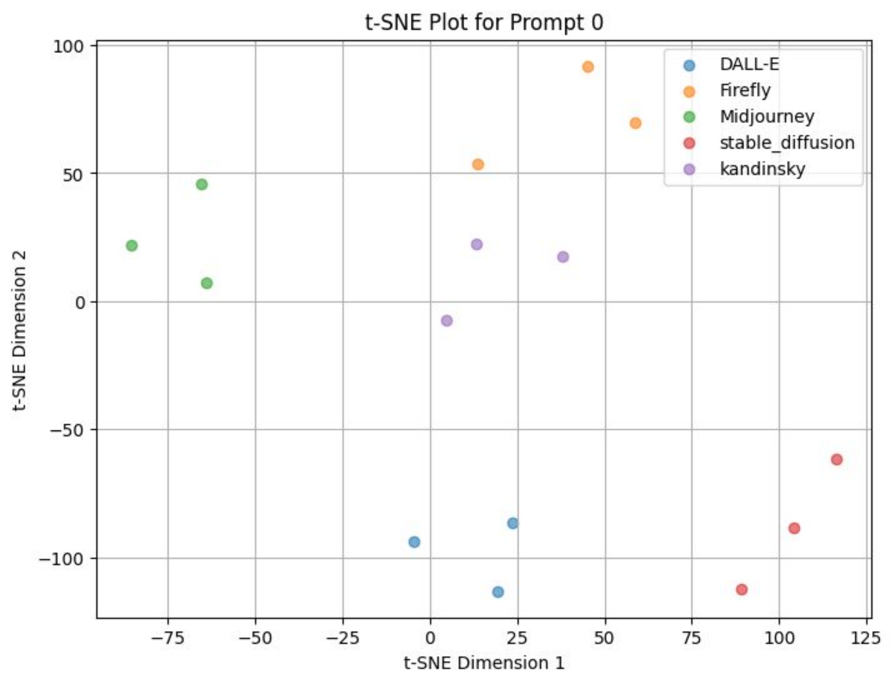
Inter-Model Comparison

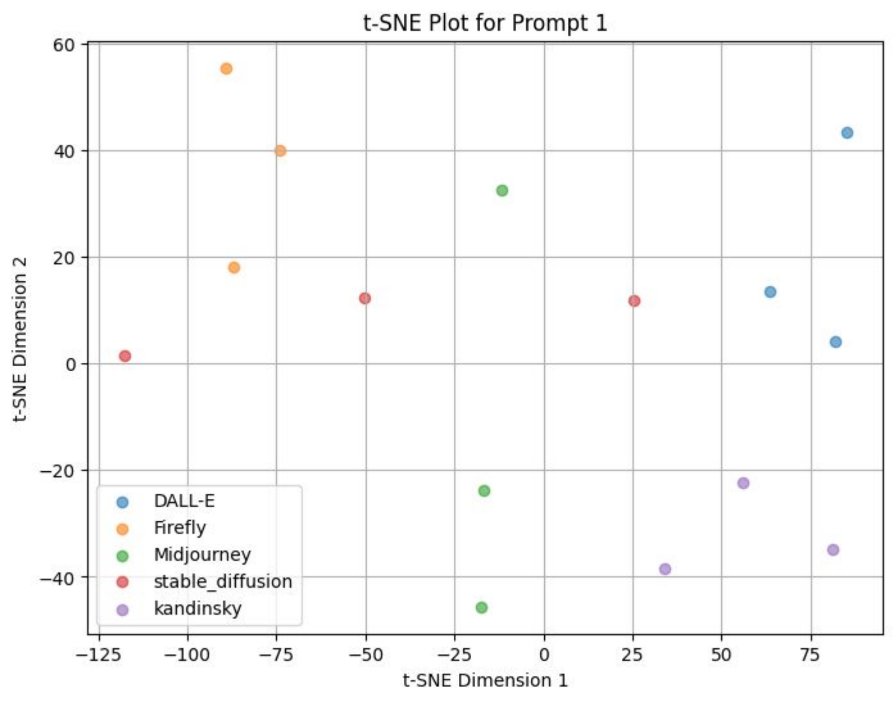


```
             model1            model2  similarity
0            DALL-E           Firefly    0.853031
1            DALL-E        Midjourney    0.885129
2            DALL-E         kandinsky    0.886584
3            DALL-E  stable_diffusion    0.815420
4           Firefly        Midjourney    0.858032
5           Firefly         kandinsky    0.869942
6           Firefly  stable_diffusion    0.832362
7        Midjourney         kandinsky    0.888418
8        Midjourney  stable_diffusion    0.819312
9  stable_diffusion         kandinsky    0.853606
```

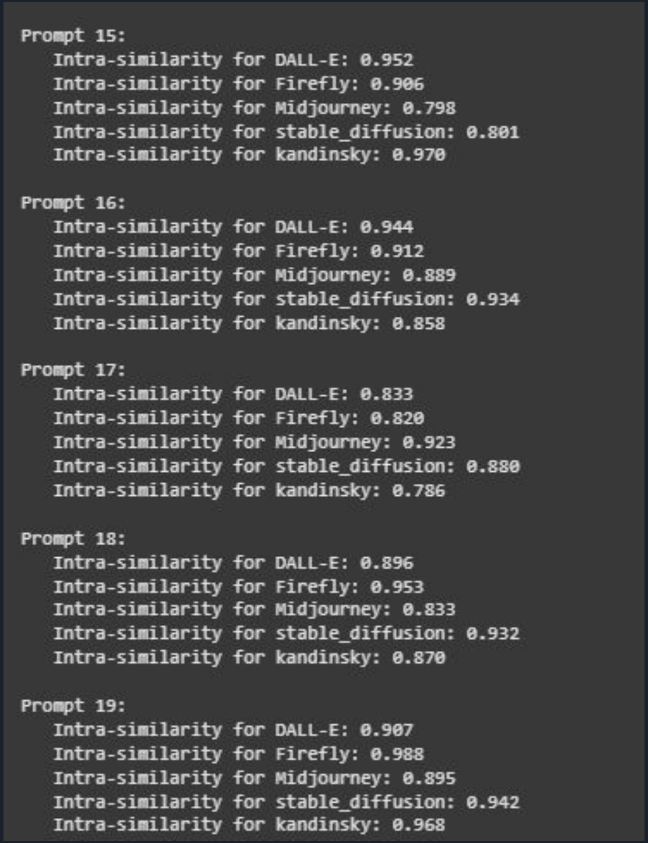


Inter-Model Comparison

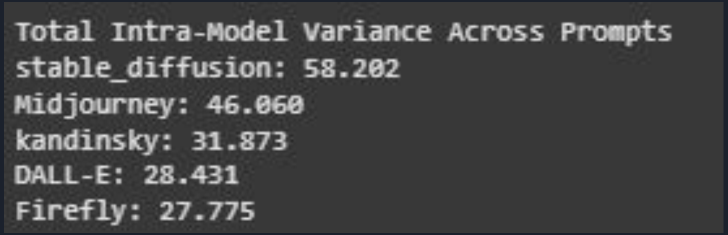




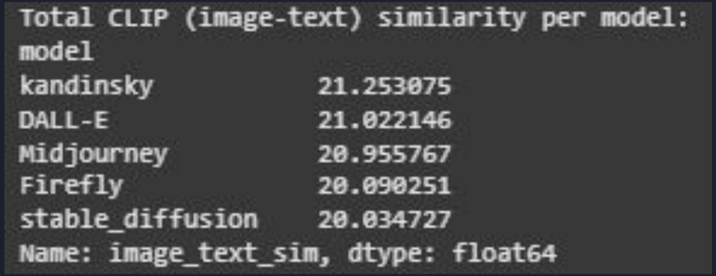
Intra-Model Comparison



```
Prompt 15:
   Intra-similarity for DALL-E: 0.952
   Intra-similarity for Firefly: 0.906
   Intra-similarity for Midjourney: 0.798
   Intra-similarity for stable_diffusion: 0.801
   Intra-similarity for kandinsky: 0.970

Prompt 16:
   Intra-similarity for DALL-E: 0.944
   Intra-similarity for Firefly: 0.912
   Intra-similarity for Midjourney: 0.889
   Intra-similarity for stable_diffusion: 0.934
   Intra-similarity for kandinsky: 0.858

Prompt 17:
   Intra-similarity for DALL-E: 0.833
   Intra-similarity for Firefly: 0.820
   Intra-similarity for Midjourney: 0.923
   Intra-similarity for stable_diffusion: 0.880
   Intra-similarity for kandinsky: 0.786

Prompt 18:
   Intra-similarity for DALL-E: 0.896
   Intra-similarity for Firefly: 0.953
   Intra-similarity for Midjourney: 0.833
   Intra-similarity for stable_diffusion: 0.932
   Intra-similarity for kandinsky: 0.870

Prompt 19:
   Intra-similarity for DALL-E: 0.907
   Intra-similarity for Firefly: 0.988
   Intra-similarity for Midjourney: 0.895
   Intra-similarity for stable_diffusion: 0.942
   Intra-similarity for kandinsky: 0.968
```

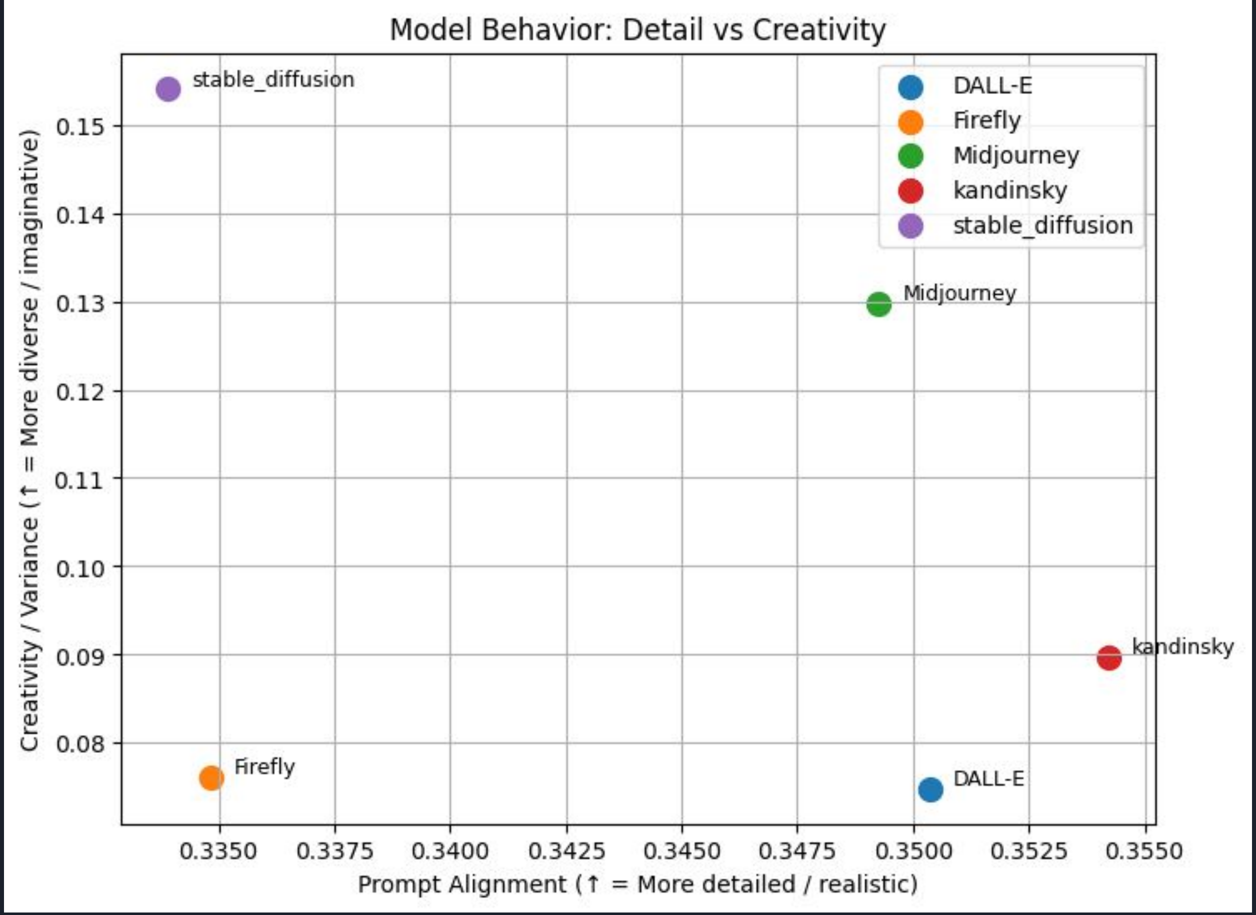


```
Total Intra-Model Variance Across Prompts
stable_diffusion: 58.202
Midjourney: 46.060
kandinsky: 31.873
DALL-E: 28.431
Firefly: 27.775
```



```
Total CLIP (image-text) similarity per model:
model
kandinsky           21.253075
DALL-E              21.022146
Midjourney          20.955767
Firefly             20.090251
stable_diffusion    20.034727
Name: image_text_sim, dtype: float64
```

Intra-Model Comparison



Conclusion

- **Semantic Agreement**: All models showed strong semantic consistency across prompts, with tight UMAP clusters indicating shared understanding of prompt meaning.

- **Prompt Dominance**: Prompt identity was the strongest driver of image embeddings across models, more so than model architecture.

- **Embedding Similarity**: Models use similar CLIP-based text/image embedding steps, but diverge during denoising and decoding, causing slight (~0.8–0.9) embedding differences.

- **Creative Variability vs Prompt Fidelity**:

  - *Stable Diffusion*: Highest variance, most creative, but least faithful to prompts.

  - *DALL-E & Kandinsky*: High prompt adherence, low variability
    - Great for precise outputs.

  - *Midjourney*: Balanced in creativity and prompt alignment.

  - *Firefly*: Most colorful, but lowest in variance and fidelity
    - Best if IP-safe datasets are a priority.
- **Practical Use Cases**: Users can choose models based on desired creativity, colorfulness, prompt faithfulness, or training data policies.

- **CLIP as Analytical Lens**: Validated CLIP as a powerful tool for comparing generative models through alignment, clustering, and consistency.

Future Work

- **Expand Dataset Scope**: Increase from 20 to 50–100 prompts across more categories to improve generalizability and robustness of findings.
- **Add Real Image Baseline**: Incorporate real-world images to better assess realism and anchor results in natural image statistics.
- **Introduce More Models**: Include additional models (e.g., Leonardo AI, Canvas, Gemini) and newer versions to capture evolving trends.
- **Use Machine Learning for Source Detection**: Train classifiers on larger datasets to automatically identify the model that generated an image.
- **Test AI Detection Without ML**: Compare unknown images to known generations using cosine similarity to infer if they're AI-generated.
- **Build a Detection & Recommendation Tool**: Develop a user-facing system to identify AI-generated content, suggest ideal models, and expose biases or hallucinations.