DIVA: Diversity Assessment in Text-to-Image Generation via Hybrid Metrics

Md Asif Bin Syed [*1] **Md Younus Ahamed** [*1]

Abstract

Generative models like Stable Diffusion, DALL·E, and Imagen have shown impressive capabilities in creating visually compelling images from textual prompts. However, not all models produce a wide variety of outputs from the same prompt. In some applications—such as creative advertising or artistic design—diverse outputs are highly valued for exploring different visual interpretations. In contrast, tasks like forensic analysis or technical illustration require high consistency to ensure reproducibility. Current diversity quantification methods, such as Bayesian frameworks and pixel-based metrics (e.g., FID, SSIM), either ignore prompt-specific variability or fail to disentangle aleatoric and epistemic factors.In this work, We present DIVA, a framework quantifying diversity through hybrid diversity metrics: mean pairwise CLIP embedding distance, feature distribution variance, and information entropy. DIVA integrates these metrics into a unified diversity score, capturing both aleatoric and epistemic uncertainty. It adapts to both diversity-expected prompts and diversity-constrained prompts. Human validation shows strong correlation between our diversity score and human judgments. This work provides a scalable solution for applications requiring reliability and transparency, from creative design to medical imaging. Github repository: https://github.com/anonymous4865/diva

1. Introduction

Diversity is a fundamental quality metric for evaluating generative image models . It measures the variety and distinctiveness of outputs when sampling multiple times from the same model. In the context of image generation, diversity encompasses the range of visual variations across generated samples, including differences in content, style, composition, and other visual attributes. Diversity metrics for generative models have evolved from simple statistical approaches to more sophisticated perceptual measures. Early pixel-based approaches calculated the coefficient of variation (CV) or standard deviation across generated samples for each pixel position, then average these values to produce an overall diversity score (Chen et al., 2023; Dubiński et al., 2022). Higher values indicate greater diversity by measuring how consistently pixel locations vary across multiple generated images.

Structural similarity measures provide another fundamental approach where the Structural Similarity Index Measure (SSIM) and its multi-scale variant MS-SSIM are widely used to assess the similarity between image pairs. When applied to diversity measurement, a lower average SSIM between randomly sampled image pairs indicates higher diversity since the images differ more significantly in their structural properties (Cha et al., 2019; Friedrich et al., 2024). Mariani et al. (Mariani et al., 2018) observed that real images were always more variable than the generated ones (lower SSIM) providing a valuable reference point for evaluating generated image diversity.

The field has advanced towards perceptual features extracted from deep neural networks. The Learned Perceptual Image Patch Similarity (LPIPS) metric has emerged as particularly valuable for evaluating generative model diversity (Zhu et al., 2017; Abbasnejad et al., 2023; Hall et al., 2023) . LPIPS computes the weighted L2 distance between deep features of image pairs, with higher average distances indicating greater diversity. Zhang et al. (Zhang et al., 2018) demonstrated that perceptual similarity appears to be an emergent property shared across deep visual representations making LPIPS robust across different architectures and training approaches. For conditional generation tasks, specialized diversity metrics have been developed. For example, in Semantically Multi-modal Image Synthesis (SMIS), separate diversity metrics evaluate variation within specific semantic regions (mean Class-Specific Diversity) versus consistency in other regions (mean Other-Classes Diversity) (Zhu et al., 2020).

Some researchers complement standard diversity metrics

[*]Equal contribution [1]West Virginia University, Morgantown, WV, USA. Correspondence to: Md Asif Bin Syed <ms00110@mix.wvu.edu>, Md Younus Ahamed <ma00087@mix.wvu.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Prompt: Create an image with the text 'Muslims in ML Workshop' at the top. Below this text, include 'ICML 2025'. The image should feature the Music City Center in the lower right and cityscapes on the left.
Diversity Score: 6.05
Average Human Diversity Rating: 2.75 (Scale: 1-5, with 5 indicating the highest diversity)
Correlation of Human Rating and Diversity Metric: 0.63
Strong Positive Correlation

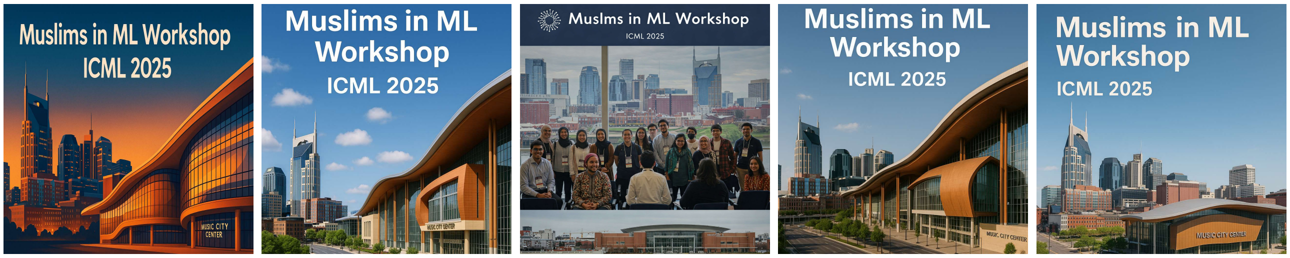

Figure 1. Visualization of diversity quantification results for a given prompt. The proposed **DIVA** framework yields a **Diversity Score of 6.05**, with an **Average Human Diversity Rating of 2.75/5** and a **Correlation of Human Rating and Diversity Metric of 0.63**.

with simpler filtering-based approaches. For instance, Duan et al. proposes a diversity rate that calculates the percentage of unique images within a generated set, where images with consistent spatial patterns are considered duplicates (Duan & Zhang, 2022). In contrast, Boutin et al. measure diversity as the standard deviation of generated samples in a learned feature space, using different feature extractors to ensure robustness (Boutin et al., 2022). More recently, Image-Image CLIP scores and Image Entropy have been employed as diversity indicators. Image-Image CLIP scores assess structural congruity between images, with lower scores between image pairs indicating higher diversity (Sun et al., 2024). Similarly, Image Entropy quantifies the average information per pixel, with higher entropy values suggesting greater diversity in information content .

While recent approaches employ Image-Image CLIP scores and Image Entropy as diversity indicators, these methods remain limited in their scope and applicability. Image-Image CLIP scores focus on structural congruity between pairs of images, which, while useful for pairwise comparisons, fail to capture holistic diversity across entire sets of outputs for a given prompt. Similarly, Image Entropy quantifies pixel-level information content but neglects semantic alignment with textual prompts, a critical factor in text-to-image generation. Our study advances this paradigm by integrates mean pairwise distance (measuring diversity via CLIP embeddings), variance (assessing output consistency), and entropy (capturing unpredictability). By combining these metrics, our approach effectively captures both aleatoric uncertainty and epistemic uncertainty. Additionally, it is validated against human judgments using crowd-sourced diversity ratings. The results revealed a strong correlation (Spearman's correlation) between our metrics and human perceptions, confirming that our methodology effectively aligns computational assessments with real-world intuitive judgments.

2. Methodology

This section describes the methodology used to assess the diversity of text-to-image models, including DALL-E 3(Betker et al., 2023), Imagen 3(Baldridge et al., 2024), and Stable Diffusion 3.5(Esser et al., 2024). The study involved generating images from a curated set of textual prompts, extracting feature representations using the CLIP model, and quantifying diversity through statistical metrics. A composite diversity score was formulated and correlated with human assessments of diversity to validate the proposed approach.

Algorithm 1 Diversity Evaluation in Text-to-Image Models

```
Input  : Prompts $P$, models $\{M_1, M_2, M_3\}$, images per prompt $N$
Output : Diversity score $U$, correlation with human ratings
Image Generation and Preprocessing
  for each prompt $p \in P$, model $M_i$ do
    Generate $N$ images using $M_i$ with standardized settings
      Extract CLIP embeddings
Diversity Computation
  for each prompt-model set do
    Compute MPD, Variance, Cross-Entropy
Normalize and compute $U = w_1 \cdot \text{MPD} + w_2 \cdot \text{Variance} + w_3 \cdot \text{Cross-Entropy}$
Human Evaluation
  Collect diversity ratings (scale 1-5), compute inter-rater agreement
Correlation Analysis
  Compute Spearman correlations between $U$ and human ratings
```

2.1. Dataset and Prompt Selection

To ensure a robust evaluation of model diversity, we selected a set of 100 textual prompts representing various semantic categories, including objects, scenes, abstract concepts, and human figures. The prompts were sourced from prior research on text-to-image generation, publicly available datasets, and manually crafted queries to ensure linguistic diversity and a broad range of possible visual outputs. Each prompt was used to generate images across the three selected text-to-image models. To minimize bias from model-specific settings, default or recommended hyperparameters were used, ensuring comparability. The number of images generated per prompt per model was fixed at 5 to maintain statistical consistency.

2.2. Image Generation and Preprocessing

The selected models were used to generate images for each prompt multiple times to capture variation in output diversity. The following preprocessing steps were applied to ensure consistency:

- The same temperature and sampling strategy (where applicable) were used across all models.
- All images were resized to a fixed resolution (1024×1024 pixels) to standardize input dimensions.
- Normalization were applied across all images.

2.3. Feature Extraction Using CLIP

To quantitatively assess image diversity, we employed the CLIP (Contrastive Language–Image Pretraining) model to extract feature representations. Specifically, we used the image encoder of CLIP to obtain high-dimensional feature embeddings for each generated image. These embeddings capture both semantic and structural properties, enabling meaningful diversity analysis. The embeddings were extracted using the pre-trained CLIP ViT-B/32 model (OpenAI, 2021), which offers robust alignment between visual and textual representations. Each image was processed through the CLIP model, and the resulting feature vectors were stored for further statistical analysis.

2.4. Diversity Measurement and Diversity Computation

To quantify diversity, we employ three statistical metrics: Mean Pairwise Distance (MPD), Variance of Embeddings, and Cross-Entropy Score. Each metric captures different aspects of diversity in generated images. The MPD measures the average cosine distance between all image embeddings within a prompt-model combination:

$$MPD = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (1 - \cos(x_i, x_j)) \tag{1}$$

where x_i and x_j are CLIP embeddings, and $\cos(x_i, x_j)$ represents their cosine similarity. The Variance of Embeddings quantifies the spread of feature representations:

$$\text{Var} = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2 \tag{2}$$

where $\bar{x}$ is the mean embedding vector. The Cross-Entropy Score assesses the unpredictability of generated images based on a probabilistic model:

$$H = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|p) \tag{3}$$

where $p(x_i|p)$ is the likelihood of an image given its prompt. A combined diversity score is computed as:

$$U = w_1 \cdot MPD + w_2 \cdot \text{Var} + w_3 \cdot H \tag{4}$$

where w_1, w_2, and w_3 represent the weights assigned to each metric component. In this study, all components are given equal weight distribution.

2.5. Human Evaluation and Correlation Analysis

To validate the computational diversity score U, a human evaluation study was conducted. Five independent evaluators assessed the diversity of generated images on a 5-point scale, where a score of 1 indicated minimal diversity (highly similar images), and a score of 5 represented high diversity (distinct and varied images). The evaluation process was blinded to prevent model bias. The final human diversity score D for each prompt-model combination was computed as the average of the two ratings:

$$D = \frac{1}{N} \sum_{i=1}^{N} d_i \tag{5}$$

where d_i represents the rating given by evaluator i. To examine the relationship between the computed diversity score U and the human-assessed diversity score D, Spearman's rank correlation coefficient ρ was calculated to assess monotonic relationships.

3. Results

This section presents the results of our diversity evaluation for text-to-image models. The analysis focuses on the computed diversity metrics, human evaluation scores, and their correlation to validate the effectiveness of the proposed diversity measure. Table 1 summarizes the diversity metrics, human evaluation scores, and correlation values for each model across two prompt conditions: **Diversity Expected** and **Diversity Not Expected**. The table includes the Mean Pairwise Distance (MPD), pixel-level variance,

Table 1. Comparison of different text-to-image models based on MPD, Variance, Entropy, overall diversity, human evaluation, and their correlation. Mean ± SD are reported for each condition.

Model	Prompt Type	MPD	Variance	Entropy	Diversity	Human Eval	Corr
DALL-E 3	Diversity Expected	0.15 ± 0.06	17.89 ± 6.02	0.59 ± 0.09	3.07 ± 1.28	2.03 ± 0.81	0.7
	Diversity Not Expected	0.08 ± 0.03	10.56 ± 3.91	0.55 ± 0.08	3.21 ± 1.08	2.2 ± 0.96	0.5
Imagen 3	Diversity Expected	0.13 ± 0.04	14.45 ± 3.72	0.60 ± 0.09	5.61 ± 1.84	2.63 ± 0.82	0.2
	Diversity Not Expected	0.11 ± 0.05	12.65 ± 5.75	0.55 ± 0.08	3.37 ± 1.18	1.15 ± 0.58	0.6
StableDiffusion 3.5	Diversity Expected	0.09 ± 0.04	9.49 ± 4.19	0.60 ± 0.09	4.54 ± 1.13	2.63 ± 0.89	0.4
	Diversity Not Expected	0.08 ± 0.03	10.64 ± 3.06	0.56 ± 0.08	3.82 ± 1.74	2.35 ± 0.90	0.5

and image entropy—each contributing to the overall computed diversity score. MPD reflects the average distance between image embeddings, variance captures pixel-wise spread across generated samples, and entropy quantifies distributional randomness across outputs. . The computed diversity score (U) was derived using the combined MPD, Variance, and Cross-Entropy metrics, while human evaluators provided diversity ratings (D) on a 5-point scale. For all models, the diversity metrics were generally higher when diversity was expected in the prompt. Imagen 3 exhibited the highest diversity score under this condition (5.61 ± 1.84), followed by Stable Diffusion 3.5 (4.54 ± 1.13) and DALL-E 3 (3.07 ± 1.28). The human evaluation scores also followed a similar pattern, with higher diversity ratings for prompts where diversity was expected. However, the alignment between computed diversity and human ratings varied across models. The correlation analysis revealed differing relationships between the computed diversity metrics and human evaluations across models. DALL-E 3 demonstrated the highest correlation ($\rho = 0.7$) when diversity was expected, suggesting that its diversity measure effectively captured the diversity perceived by human evaluators. However, this correlation dropped to 0.5 when diversity was not explicitly expected. Imagen 3 exhibited an inverse trend, with a weak correlation ($\rho = 0.2$) for diversity-expected prompts but a stronger correlation ($\rho = 0.6$) for diversity-not-expected prompts. This indicates that the computed diversity measure was less effective in capturing diversity when variation was anticipated but aligned better when less diversity was expected. Stable Diffusion 3.5 displayed moderate correlations ($\rho = 0.4$ and 0.5), indicating a relatively stable relationship between computed and human-assessed diversity scores across both prompt types. Overall, DALL-E 3 demonstrated the most consistent alignment between computed diversity and human perception in diversity-expected prompts. Imagen 3 exhibited the highest absolute diversity scores but weaker correlations with human evaluations in diversity-expected cases. Stable Diffusion 3.5 maintained a more balanced performance across conditions. These results suggest that while diversity metrics can effectively quantify diversity in generated images, their alignment with human perception varies across models and prompt types. Further optimization of weighting parameters in the diversity formula could enhance its predictive capability across diverse generative models.

4. Challenges and Future Scope

Despite its contributions, DIVA faces limitations: (1) Feature-space bias from CLIP's pretraining data skews diversity assessments for culturally or semantically niche prompts; (2) Human evaluation scalability remains constrained by the labor-intensive nature of crowdsourced annotations and inter-rater variability; (3) Static metric weighting fails to adapt to prompts with varying semantic granularity, limiting dynamic diversity disentanglement (aleatoric vs. epistemic); (4) Computational overhead persists in large-N sampling regimes despite subsampling optimizations. Future work will prioritize adversarial debiasing of CLIP embeddings, self-supervised quality estimation models (e.g., diffusion critic networks) to proxy human judgments, and attention-based dynamic weighting conditioned on prompt ambiguity embeddings. Extending the framework to multimodal joint diversity spaces—via cross-modal attention mechanisms between text and image latent variables—could enable end-to-end diversityx‘-aware generation.

References

Abbasnejad, I., Zambetta, F., Salim, F., Wiley, T., Chan, J., Gallagher, R., and Abbasnejad, E. Scone-gan: Semantic contrastive learning-based generative adversarial network for an end-to-end image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1111–1120, 2023.

Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Castrejon, L., Chan, K., Chen, Y., Dieleman, S., Du, Y., et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8, 2023.

Boutin, V., Singhal, L., Thomas, X., and Serre, T. Diversity vs. recognizability: Human-like generalization in one-shot generative models. *Advances in Neural Information Processing Systems*, 35:20933–20946, 2022.

Cha, M., Gwon, Y. L., and Kung, H. Adversarial learning of semantic relevance in text to image synthesis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3272–3279, 2019.

Chen, J., Xu, Q., Kang, Q., and Zhou, M. Mogan: Morphologic-structure-aware generative learning from a single image. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(4):2021–2033, 2023.

Duan, Y. and Zhang, J. A novel ai-based visual stimuli generation approach for environment concept design. *Computational Intelligence and Neuroscience*, 2022(1):8015492, 2022.

Dubiński, J., Deja, K., Wenzel, S., Rokita, P., and Trzcinski, T. Selectively increasing the diversity of gan-generated samples. In *International Conference on Neural Information Processing*, pp. 260–270. Springer, 2022.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Friedrich, P., Frisch, Y., and Cattin, P. C. Deep generative models for 3d medical image synthesis. *arXiv preprint arXiv:2410.17664*, 2024.

Hall, M., Ross, C., Williams, A., Carion, N., Drozdzal, M., and Soriano, A. R. Dig in: Evaluating disparities in image generations with indicators for geographic diversity. *arXiv preprint arXiv:2308.06198*, 2023.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

OpenAI. Clip: Connecting text and images. https://openai.com/index/clip/, January 2021. URL https://openai.com/index/clip/. Accessed: 2025-03-14.

Sun, H., Xia, B., Chang, Y., and Wang, X. Generalizing alignment paradigm of text-to-image generation with preferences through f-divergence minimization. *arXiv preprint arXiv:2409.09774*, 2024.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

Zhu, Z., Xu, Z., You, A., and Bai, X. Semantically multi-modal image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5467–5476, 2020.