WHERE DO PIXELS HIDE THEIR SECRETS? UNCOVERING MODEL INTERNALS FOR CREATIVE TEXT-TO-IMAGE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) models can produce visually convincing images while still failing to correctly realize the concepts and relationships expressed in a prompt. This challenge is particularly difficult for creative inputs such as poems and stories, for which reference images are typically unavailable. Existing reference-free metrics primarily evaluate the relationship between the input text and the final generated image, while recent interpretability studies suggest that intermediate representations contain information about the spatial organization of generated content during the denoising process. We ask whether such model-internal signals can instead be used to evaluate generation quality. We introduce **MIRAGE (Model-Internal Representation Analysis for Generation Evaluation)**, a reference-free metric that evaluates spatial properties of internal representations during image generation. **MIRAGE** extracts image-token representations from intermediate diffusion-transformer blocks and decomposes them into sparse latent features. For each generation state, it identifies the dominant sparse feature and derives two complementary spatial signals: cross-seed spatial consistency, which measures the stability of the feature's spatial centroid across stochastic generations, and activation concentration, which measures how spatially localized the feature's activation is around its centroid. These signals are aggregated into an interpretable score for each generated image. We evaluate **MIRAGE** on a curated collection of 6,000 poems and stories generated using three diffusion-transformer T2I models. Across human evaluations of semantic fidelity, attribute presence, compositional correctness, and narrative fidelity, **MIRAGE** shows agreement with human judgments and provides a complementary signal to existing reference-free evaluation metrics. These results suggest that model-internal spatial representations can provide useful signals for reference-free evaluation of creative T2I generation.

[1]

1 INTRODUCTION

Evaluating text-to-image (T2I) generation becomes particularly challenging when the input is a creative text such as a poem or a story. Unlike short object-centric prompts, creative prompts may contain multiple concepts, abstract or metaphorical elements, and spatial relationships between them. Recent studies have shown that current T2I models still struggle with complex prompts involving multiple objects, attributes, and relationships, making fine-grained evaluation challenging Huang et al. (2023; 2025a). A generated image must therefore not only be visually aligned with the input text, but also preserve important concepts and their relationships in the generated composition. Existing evaluation metrics such as CLIP-based similarity, BLIP/VQA-based evaluation, and image-quality metrics primarily assess the final generated artifact, either by measuring its alignment with the input text or by comparing it with a reference image Radford et al. (2021); Li et al. (2022); Lin et al. (2024). However, creative T2I datasets containing poems and stories rarely provide corresponding ground-truth images, making reference-based evaluation difficult. Moreover, text-image

[1] All data, code and checkpoints are made available at: `https://anonymous.4open.science/r/MIRAGE-ICLR/`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

similarity alone may not fully capture whether important concepts are realized correctly or whether their relationships are preserved. Recent evaluations of T2I metrics have similarly shown that different metrics capture different aspects of generation quality and that their agreement with human judgments can vary across evaluation settings Wiles et al. (2025). For example, a generated image may remain broadly aligned with a prompt while failing to represent an important concept, placing a concept inconsistently across generations, or weakly associating a generated region with the corresponding textual concept. Recent interpretability studies provide evidence that the internal computations of diffusion models contain information about how textual concepts are realized during generation. Methods based on diffusion attention and intermediate representations have been used to localize textual concepts and analyze their relationships with generated image regions Tang et al. (2023); Chefer et al. (2024); Kong et al. (2024); Helbling et al. (2025). More recently, sparse autoencoders have been applied to diffusion models to decompose intermediate representations into interpretable features and study how these features evolve during generation Surkov et al. (2025); Huang et al. (2025b). These studies establish that internal representations and attention can provide meaningful information about concept realization. However, they primarily use such signals for interpretability, localization, or analysis rather than as a quantitative signal for evaluating the quality of an individual generated image.

These limitations motivate us to investigate whether model-internal representations can provide useful signals for reference-free evaluation. We hypothesize that successful generations exhibit spatially consistent and concentrated internal representations across stochastic generations, providing an additional signal beyond final-image metrics. This leads to three research questions:

1. **RQ1:** Can internal representations of the generation process provide useful signals for evaluating generated images?
2. **RQ2:** When a concept is consistently realized in the generated image, does its internal representation exhibit stable spatial behavior across denoising steps and stochastic generations?
3. **RQ3:** Can these internal representation dynamics and attention signals provide a reliable reference-free signal for evaluating creative T2I generation?

To investigate these questions, we propose **MIRAGE (Model-Internal Representation Analysis for Generation Evaluation)**, a reference-free metric that uses sparse image-token representations from intermediate diffusion-transformer blocks to measure cross-seed spatial consistency and activation concentration.

Our contributions are summarized as follows:

1. We propose **MIRAGE**, a model-internal, reference-free metric for evaluating creative text-to-image generation.
2. We introduce a block-level analysis of concept-specific internal representations and text-image attention behavior across three diffusion-transformer-based T2I models.
3. We conduct a large-scale evaluation on a collection of poems and stories and show that **MIRAGE** exhibits agreement with human judgments of generated-image quality.

2 Problem Formulation

Let P denote a creative text prompt and $\mathcal{C}(P) = \{c_1, \ldots, c_K\}$ denote its salient concepts. Let G be a diffusion-transformer-based text-to-image model with L selected transformer blocks and T denoising steps. For a stochastic seed S, the generated image is

$$I_S = G(P, S). \tag{1}$$

We seek a reference-free evaluation function

$$M(P, G), \tag{2}$$

that quantifies the spatial organization and consistency of model-internal representations without requiring a ground-truth image.

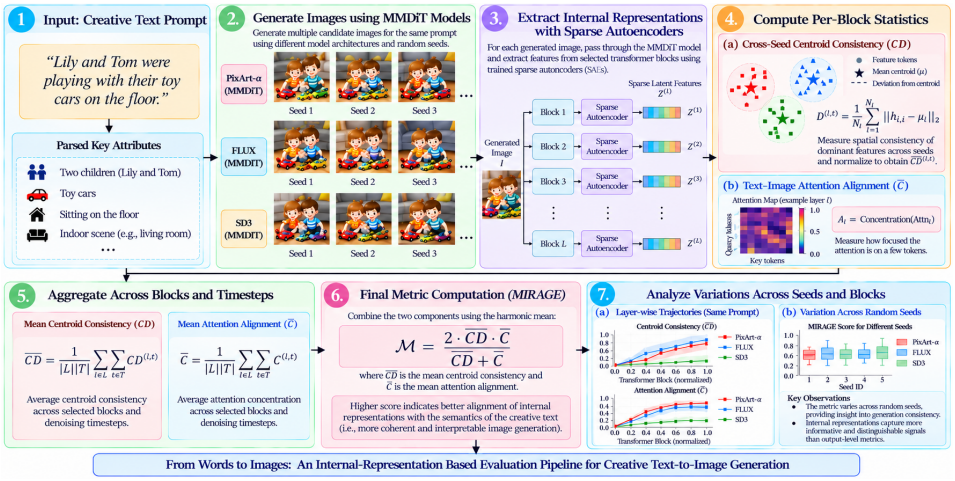

Figure 1: Overview of proposed textbfMIRAGE framework

For block l and denoising step t, let

$$H^{(l,t)}(S) \in \mathbb{R}^{N \times d_l} \tag{3}$$

denote the image-token representations, where N is the number of image tokens and d_l is the hidden dimension. A sparse autoencoder F_l with latent dimension D is trained for each selected block, with

$$Z^{(l,t)}(S) = F_l\Big(H^{(l,t)}(S)\Big) \in \mathbb{R}^{N \times D}. \tag{4}$$

For each (P, S, l, t), the dominant sparse feature is selected as

$$j^{(l,t)}(S) = \arg\max_{j \in [D]} \frac{1}{N} \sum_{n=1}^{N} Z^{(l,t)}_{n,j}(S). \tag{5}$$

Its token activations are reshaped into a spatial heatmap $A^{(l,t)}(S)$, from which the spatial centroid is computed as

$$\mu^{(l,t)}(S) = \frac{\sum_{u,v}(u,v)A^{(l,t)}(S)[u,v]}{\sum_{u,v} A^{(l,t)}(S)[u,v] + \epsilon}. \tag{6}$$

Given seeds $\mathcal{S} = \{S_1, \ldots, S_m\}$, the cross-seed spatial deviation at block l and timestep t is

$$D^{(l,t)} = \frac{1}{m} \sum_{r=1}^{m} \left\| \mu^{(l,t)}(S_r) - \bar{\mu}^{(l,t)} \right\|_2, \qquad \bar{\mu}^{(l,t)} = \frac{1}{m} \sum_{r=1}^{m} \mu^{(l,t)}(S_r). \tag{7}$$

Lower $D^{(l,t)}$ indicates greater spatial consistency of the internal representation across stochastic generations.

To account for the spatial extent of the representation, we compute its activation concentration $C^{(l,t)}(S)$ from the normalized weighted spatial variance of the heatmap. The corresponding spatial radius is

$$R^{(l,t)}(S) = \sqrt{\big(1 - C^{(l,t)}(S)\big)\, V_{\max}}, \qquad V_{\max} = \frac{W^2 + H^2}{4}, \tag{8}$$

where W and H are the heatmap width and height. The normalized cross-seed deviation is therefore

$$D^{(l,t)}_{\text{norm}} = \frac{D^{(l,t)}}{\frac{1}{m}\sum_{r=1}^{m} R^{(l,t)}(S_r) + \epsilon}. \tag{9}$$

We convert the normalized deviation into a higher-is-better cross-seed consistency score:

$$CD^{(l,t)} = 1 - D^{(l,t)}_{\text{norm}}. \tag{10}$$

The resulting spatial consistency and activation-concentration statistics are aggregated across the selected blocks and denoising steps to obtain the model-internal reference-free evaluation score $M(P, G)$.

3 DATA AND MODELS

Data: To analyse the working of our proposed metric **MIRAGE**, we experimented on a curated dataset of **6000** poems and stories. We drew poem samples from the benchmark dataset **PoemSum Mahbub et al. (2023)**, and story samples from the benchmark datasets **TinyStories Eldan & Li (2023)** and **VIST Ting-Hao et al. (2016)**. For each textual input, images are generated independently using multiple stochastic seeds. The resulting generations provide the basis for measuring the consistency of internal representations under stochastic variation. We split the data into a **70:30** split, resulting in 4200 samples for SAE training and 1800 for testing.

Models: To stabilize the proposed metric **MIRAGE** across different transformer-based diffusion models, we experimented across three model families, namely **Stable Diffusion V3** Esser et al. (2024), **Flux.1 Schenell** and **Pixart-alpha** Xu et al. (2023) models. These models provide diverse transformer architectures for evaluating whether the proposed internal signals remain useful across different model designs. For each model, we extract intermediate image-token representations from selected transformer blocks during the denoising process.

Block Selection. We select transformer blocks to provide coverage across network depth without requiring activation extraction from every block. For SD3 and PixArt, blocks $\{4, 8, 12, 16, 20\}$ are used to obtain approximately uniform early-to-late coverage. For FLUX, which contains distinct double-stream and single-stream processing stages, we analyze double-stream early, double-stream late, single-stream early, and single-stream late blocks. The selection is fixed independently of human-evaluation scores and is intended to provide representative architectural coverage rather than optimize layer selection. Detailed block-wise results are provided in Appendix D.

4 METHODOLOGY

4.1 OVERVIEW

MIRAGE evaluates generated images using internal representations captured during denoising. It decomposes image-token activations with sparse autoencoders, measures their spatial consistency across stochastic seeds and their activation concentration, and combines these signals into a reference-free score. Figure 1 provides the overall pipeline of our proposed metric.

4.2 INTERNAL REPRESENTATION EXTRACTION

Let G denote a diffusion-transformer-based text-to-image model. Given a prompt P and seed S, forward hooks capture intermediate hidden states from selected transformer blocks at the denoising steps of interest. For block l and timestep t, the captured representation is

$$H^{(l,t)}(S) \in \mathbb{R}^{B\times N\times d_l}, \tag{11}$$

where B is the batch size, N is the sequence length, and d_l is the block's hidden dimension. The extractor isolates the image-token representations from each block output; for classifier-free-guidance models, the conditional representation is retained. Blocks are selected according to each model's architecture: SD3 and PixArt use sequential transformer blocks, while FLUX provides double-stream and single-stream blocks. After extraction, the same **MIRAGE** processing is applied to the resulting image-token representations across all three models.

4.3 SPARSE FEATURE DECOMPOSITION

To obtain a sparse representation of the intermediate image-token activations, **MIRAGE** trains one sparse autoencoder for each selected transformer block. Let

$$F_l : \mathbb{R}^{d_l} \to \mathbb{R}^D \tag{12}$$

denote the SAE associated with block l, where D is the dimension of its latent space.

For an image-token representation $H^{(l,t)}(S)$, the corresponding sparse code is

$$Z^{(l,t)}(S) = F_l\left(H^{(l,t)}(S)\right) \in \mathbb{R}^{N\times D}. \tag{13}$$

The SAE consists of a linear encoder followed by a ReLU activation and a linear decoder. Its training objective is

$$\mathcal{L}_{\text{SAE}} = \left\| \widehat{H} - H \right\|_2^2 + \lambda \left\| Z \right\|_1 , \tag{14}$$

where $\widehat{H}$ is the reconstructed representation and λ controls the sparsity penalty. The decoder columns are constrained to unit norm after optimization updates to prevent trivial rescaling of the decoder from reducing the effective sparsity penalty.

4.4 Sparse Feature-Based Spatial Localization

For each prompt, seed, transformer block, and evaluated denoising step, the SAE produces a sparse activation matrix $Z^{(l,t)}(S)$. Rather than selecting a latent using an explicit text-token-to-feature correspondence, **MIRAGE** selects the dominant sparse feature according to its mean activation over the image tokens:

$$j^{(l,t)}(S) = \arg \max_{j \in \{1,\dots,D\}} \frac{1}{N} \sum_{n=1}^{N} Z_{n,j}^{(l,t)}(S). \tag{15}$$

The token-level activations of the selected feature are then reshaped into the spatial arrangement of the image tokens and bilinearly upsampled to a fixed 64×64 resolution, producing a spatial activation map

$$A^{(l,t)}(S) \in \mathbb{R}^{64 \times 64}. \tag{16}$$

The map represents the spatial distribution of the dominant sparse feature in the image-token representation for the given generation state.

4.5 Spatial Centroid and Activation Concentration

Given the spatial activation map $A^{(l,t)}(S)$, we compute its non-negative weighted spatial centroid. Let (u, v) denote a spatial location in the 64×64 map. The centroid is defined as

$$\mu^{(l,t)}(S) = \frac{\sum_{u,v} (u, v) A^{(l,t)}(S)[u, v]}{\sum_{u,v} A^{(l,t)}(S)[u, v] + \epsilon}. \tag{17}$$

To quantify the spatial concentration of the activation, we compute the weighted second moment around the centroid:

$$V^{(l,t)}(S) = \frac{\sum_{u,v} \left\| (u, v) - \mu^{(l,t)}(S) \right\|_2^2 A^{(l,t)}(S)[u, v]}{\sum_{u,v} A^{(l,t)}(S)[u, v]}. \tag{18}$$

Let

$$V_{\max} = \frac{W^2 + H^2}{4}, \tag{19}$$

where $H = W = 64$. The activation concentration is then defined as

$$C^{(l,t)}(S) = \max \left(0, 1 - \frac{V^{(l,t)}(S)}{V_{\max}} \right). \tag{20}$$

Higher values of $C^{(l,t)}(S)$ indicate that the activation mass is more spatially concentrated around its centroid.

4.6 CROSS-SEED SPATIAL CONSISTENCY

MIRAGE measures the stability of the spatial organization of the dominant sparse feature under stochastic variation. For a fixed prompt, we generate images using a set of stochastic seeds

$$\mathcal{S} = \{S_1, S_2, \ldots, S_m\}. \tag{21}$$

For a fixed block l and denoising step t, let

$$\mu_r^{(l,t)} = \mu^{(l,t)}(S_r) \tag{22}$$

denote the spatial centroid obtained from seed S_r. The cross-seed mean centroid is

$$\bar{\mu}^{(l,t)} = \frac{1}{m} \sum_{r=1}^{m} \mu_r^{(l,t)}. \tag{23}$$

We quantify cross-seed spatial dispersion using the mean Euclidean distance of the individual centroids from their cross-seed mean:

$$D^{(l,t)} = \frac{1}{m} \sum_{r=1}^{m} \left\| \mu_r^{(l,t)} - \bar{\mu}^{(l,t)} \right\|_2. \tag{24}$$

Lower values of $D^{(l,t)}$ indicate greater spatial consistency across stochastic generations.

To account for differences in the spatial extent of the activation, we estimate a characteristic radius from the concentration score:

$$R^{(l,t)}(S_r) = \sqrt{\left(1 - C^{(l,t)}(S_r)\right) V_{\max}}. \tag{25}$$

The mean radius across seeds is

$$\bar{R}^{(l,t)} = \frac{1}{m} \sum_{r=1}^{m} R^{(l,t)}(S_r). \tag{26}$$

We then normalize the centroid dispersion by the mean spatial radius:

$$D_{\text{norm}}^{(l,t)} = \frac{D^{(l,t)}}{\bar{R}^{(l,t)} + \epsilon}. \tag{27}$$

Finally, we express the result in a higher-is-better form:

$$CD^{(l,t)} = 1 - D_{\text{norm}}^{(l,t)}. \tag{28}$$

Higher $CD^{(l,t)}$ therefore indicates greater cross-seed spatial consistency relative to the spatial extent of the activation.

4.7 AGGREGATION ACROSS BLOCKS AND TIMESTEPS

For each prompt, the centroid-consistency component is obtained by averaging the valid $CD^{(l,t)}$ values across the selected transformer blocks and evaluated denoising steps:

$$CD = \frac{1}{|\mathcal{V}|} \sum_{(l,t) \in \mathcal{V}} CD^{(l,t)}, \tag{29}$$

where $\mathcal{V}$ denotes the set of valid block–timestep evaluations for the prompt.

Similarly, the activation-concentration component is obtained by averaging the corresponding concentration values:

$$\overline{C} = \frac{1}{|\mathcal{V}|} \sum_{(l,t)\in\mathcal{V}} C^{(l,t)}. \tag{30}$$

Both components are defined such that larger values indicate stronger spatial consistency or concentration.

4.8 Activation Concentration Component

The second **MIRAGE** component captures the spatial concentration of the dominant sparse feature rather than its cross-seed stability. For each prompt, block, timestep, and seed, the concentration score $C^{(l,t)}(S)$ is computed from the weighted spatial variance of the corresponding SAE activation map. We aggregate these values over the valid block–timestep evaluations to obtain

$$\overline{C} = \frac{1}{|\mathcal{V}|} \sum_{(l,t)\in\mathcal{V}} C^{(l,t)}. \tag{31}$$

Higher $\overline{C}$ indicates that the dominant sparse feature is more spatially concentrated in the generated representation.

4.9 **MIRAGE** Score

MIRAGE combines two complementary properties of the model-internal activation: cross-seed spatial consistency and spatial activation concentration. Let CD denote the aggregated centroid-consistency score and $\overline{C}$ denote the aggregated activation-concentration score.

Because the two components capture complementary properties that are both required for a strong internal spatial signal, we use their harmonic mean. Unlike the arithmetic mean, the harmonic mean penalizes an imbalance between the two components and therefore prevents a high value of one component from fully compensating for a low value of the other:

$$\boxed{\text{MIRAGE} = \frac{2\,CD\,\overline{C}}{CD + \overline{C}}} \tag{32}$$

Both components are higher-is-better quantities. CD measures whether the spatial organization of the dominant sparse feature is consistent across stochastic generations, whereas $\overline{C}$ measures how spatially concentrated the feature activation is within the generated image. The harmonic formulation therefore assigns a high MIRAGE score only when both spatial consistency and activation concentration are simultaneously high, providing a conservative combination of the two complementary internal signals.

5 Results

5.1 MIRAGE Components

We first examine the two internal signals underlying **MIRAGE**: cross-seed spatial consistency and activation concentration. For a fair comparison across architectures, we evaluate an identical set of 193 prompts for each of FLUX, SD3, and PixArt. Thus, all model-level statistics reported in this section are computed from a balanced set of 193 prompts per model.

Table 1 summarizes the two components. The mean cross-seed consistency (CD) is 0.9344, 0.9543, and 0.9454 for FLUX, SD3, and PixArt, respectively. The corresponding mean activation concentration is 0.6792, 0.6827, and 0.6659. The consistently high CD values indicate that the spatial organization of the dominant sparse activation exhibits relatively small variation across stochastic

Table 1: **MIRAGE** components and final score on a balanced set of 193 prompts per model. CD denotes cross-seed centroid consistency, C denotes activation concentration, and **MIRAGE** is their equal-weight combination.

Model	$CD \uparrow$	$C \uparrow$	**MIRAGE** $\uparrow$
FLUX	0.9344	0.6792	0.7866
SD3	0.9543	0.6827	0.7959
PixArt	0.9454	0.6659	0.7814

generations. In contrast, activation concentration measures how spatially localized this activation is within an individual generation.

The two components exhibit distinct behavior across transformer blocks. For example, in FLUX, the early double-stream blocks achieve the highest centroid consistency ($CD = 0.9539$), whereas the late single-stream blocks exhibit the highest activation concentration ($C = 0.7088$). This difference indicates that cross-seed spatial stability and within-generation spatial localization capture complementary properties of the internal representation. Detailed block-wise results are provided in Appendix D.2.1.

5.2 Overall MIRAGE Score

On the balanced evaluation set, the resulting mean **MIRAGE** scores are 0.7866, 0.7959, and 0.7814 for FLUX, SD3, and PixArt, respectively. The scores are computed consistently from the same two model-internal spatial signals across the three diffusion-transformer architectures. The similar score ranges indicate that the proposed metric can be applied across different architectures while preserving the same interpretation of the resulting score. These values are intended to characterize the internal spatial properties captured by **MIRAGE**, rather than to rank the underlying image generators.

5.3 Complementarity of **MIRAGE** Components

The behavior of the two components across transformer blocks further illustrates their complementarity. For FLUX, the early double-stream blocks obtain the highest CD (0.9539), whereas the late single-stream blocks obtain the highest activation concentration (0.7088). Consequently, a representation can exhibit strong cross-seed spatial stability without being maximally concentrated, and vice versa. This distinction supports combining the two signals rather than relying on either statistic alone.

5.4 Cross-Architecture Evaluation

We combine the two higher-is-better components using equal weighting: The resulting mean **MIRAGE** scores are 0.7866, 0.7959, and 0.7814 for **FLUX, SD3, and PixArt**, respectively. The comparable score ranges across the three architectures show that the proposed internal statistics can be extracted and combined consistently across distinct diffusion-transformer architectures. We do not interpret these differences as a ranking of generator quality; rather, this experiment tests the applicability of the proposed representation-level metric across architectures.

5.5 Correlation with Human Judgments

We next investigate whether **MIRAGE** is associated with human judgments of creative text-to-image generation. Human annotation is costly; therefore, we construct a balanced evaluation test set of 579 prompt–model pairs across FLUX, SD3, and PixArt. Human judgments are aggregated across four dimensions: **semantic fidelity, attribute presence, compositional correctness, and narrative fidelity.**

We compare **MIRAGE** with widely used reference-free automated evaluation metrics. Table 2 reports the Pearson correlation between each metric and the aggregated human judgments. **MIRAGE** obtains a correlation of $r = 0.280$ on the complete evaluation set. The corresponding model-specific

Table 2: Pearson correlation between reference-free automated metrics and aggregated human judgments on the evaluation set of 579 prompt–model pairs.

Metric	All (579)	SD3	PixArt	FLUX
MIRAGE	**0.280**	**0.299**	**0.165**	**0.351**
CLIPScore	0.155	0.149	0.017	0.254
BLIP-2	0.244	0.289	0.193	0.263
VQAScore	0.154	0.068	0.140	0.245

correlations are $r = 0.299$ for SD3, $r = 0.165$ for PixArt, and $r = 0.351$ for FLUX. Among the output-level metrics, BLIP-2 obtains $r = 0.244$, followed by CLIPScore ($r = 0.155$) and VQAScore ($r = 0.154$). These results show that the internal signals captured by **MIRAGE** are associated with human assessment of creative generations and provide a complementary perspective to metrics computed directly from the generated image.

5.6 Discussion

The human evaluation results indicate that **MIRAGE** captures information that is relevant to human assessment of creative text-to-image generations. On the evaluated set, **MIRAGE** achieves a Pearson correlation of $r = 0.280$ with aggregated human judgments, while the corresponding correlations for CLIPScore, BLIP-2, and VQAScore are 0.155, 0.244, and 0.154, respectively. The model-specific correlations also vary across generators, with **MIRAGE** obtaining 0.299, 0.165, and 0.351 for SD3, PixArt, and FLUX, respectively. This variation indicates that the relationship between internal spatial signals and human judgments depends on the underlying generator, highlighting the complementary nature of model-internal evaluation signals.

Importantly, these results should be viewed as evidence that model-internal spatial information can provide a complementary signal for reference-free evaluation, rather than as evidence that **MIRAGE** fully captures human judgments of creative quality. The moderate correlations also indicate that semantic and narrative properties of poems and stories may not be completely represented by the spatial statistics considered by the metric. Extending the evaluation to additional architectures, larger human-annotated datasets, and other forms of internal representation would therefore be useful for assessing the broader applicability of the approach.

6 Conclusion

We presented **MIRAGE**, a reference-free evaluation metric for creative text-to-image generation that derives spatial statistics from model-internal representations and combines cross-seed spatial consistency with activation concentration. Experiments on creative prompts across SD3, PixArt, and FLUX, together with human evaluation, show that these internal signals are associated with human judgments and provide a complementary perspective to existing output-level metrics.

In its present form, **MIRAGE** has been experimentally evaluated only on **Diffusion-transformer architectures**. The current study does not extend the method or validate its behavior on **U-Net-based text-to-image models**. Consequently, claims about the general applicability of MIRAGE beyond the evaluated diffusion transformer architectures are not established by the present experiments. Extending the framework to U-Net-based generators and validating it across a broader range of architectures constitute important directions for future work.

References

Hila Chefer, Oran Lang, Mor Geva, Volodymyr Polosukhin, Assaf Shocher, michal Irani, Inbar Mosseri, and Lior Wolf. The hidden language of diffusion models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 52637–52669, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e9000ecb86d45c442a1d38fae68dd8fb-Paper-Conference.pdf.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion transformers learn highly interpretable features. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22946–22963. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/helbling25a.html.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023. URL https://arxiv.org/abs/2303.11897.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation, 2025a. URL https://arxiv.org/abs/2307.06350.

Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Fu-Yun Wang, Yuchi Wang, Renrui Zhang, Peng Gao, and Hongsheng Li. Tide : Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation, 2025b. URL https://arxiv.org/abs/2503.07050.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2305.01569.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 55279–55292, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f37265d7493377170a3b4ba91823119a-Paper-Conference.pdf.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. URL https://arxiv.org/abs/2404.01291.

Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14878–14886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.920. URL https://aclanthology.org/2023.emnlp-main.920/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 99956–100027. Curran Associates, Inc., 2025. doi: 10.52202/085713-3342. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/90291c8d685db65a8a1e32fd2b66f8bb-Paper-Conference.pdf.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.310. URL https://aclanthology.org/2023.acl-long.310/.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 166943–166986. Curran Associates, Inc., 2025. doi: 10.52202/085713-5566. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f3d1dab3b68abb7bb2af5cae28def3b0-Paper-Conference.pdf.

Ting-Hao, Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling, 2016. URL https://arxiv.org/abs/1604.03968.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Chris Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings, 2025. URL https://arxiv.org/abs/2404.16820.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A LLM USAGE STATEMENT

We used large language models only as general-purpose writing assistants, to proofread and correct grammatical errors in this manuscript.

B RELATED WORK

Text-to-Image Evaluation. Text-to-image evaluation has traditionally focused on image quality, text–image alignment, and compositional faithfulness. Embedding-based metrics such as CLIPScore Radford et al. (2021) and BLIPScore Li et al. (2022) measure correspondence between generated images and text, while learned preference models such as ImageReward Xu et al. (2023), PickScore Kirstain et al. (2023), and HPSv2 Wu et al. (2023) incorporate human preference signals. More fine-grained evaluators decompose prompts into objects, attributes, and relations using question answering or structured representations. TIFA Hu et al. (2023) evaluates text–image faithfulness through automatically generated visual questions, while T2I-CompBench Huang et al. (2023) and

its extension Huang et al. (2025a) target attribute binding, spatial and non-spatial relations, and complex compositions. VQAScore Lin et al. (2024) similarly formulates evaluation as visual question answering. These approaches provide increasingly fine-grained evaluation of the generated artifact, but remain primarily output-oriented and generally require either a reference image, an external vision-language model, or prompt-derived questions to assess the final image. Recent studies further show that different automatic metrics can exhibit substantially different agreement with human judgments across evaluation settings Wiles et al. (2025).

Interpretability of Diffusion Models. A complementary line of work investigates the internal computation of diffusion models to understand how textual concepts are realized during generation. DAAM Tang et al. (2023) aggregates cross-attention maps in Stable Diffusion to attribute generated image regions to textual tokens. The Hidden Language of Diffusion Models Chefer et al. (2024) and Interpretable Diffusion Kong et al. (2024) further analyze internal representations and information flow to study concept localization and word–region relationships. More recently, ConceptAttention Helbling et al. (2025) demonstrates that attention representations in diffusion transformers provide highly interpretable concept localization. Sparse autoencoders (SAEs) have also been introduced as a means of decomposing distributed diffusion representations into more interpretable features Surkov et al. (2025); Tinaz et al. (2025). TIDE Huang et al. (2025b) extends this direction by explicitly studying temporally varying SAE features across denoising steps in diffusion transformers. Collectively, these works establish that intermediate representations and attention contain useful information about concept realization, localization, and the evolution of generation.

Our Novelty. Unlike DAAM Tang et al. (2023), which uses cross-attention for text-to-image attribution, Hidden Language and Interpretable Diffusion Chefer et al. (2024) , which study semantic information in diffusion representations, and ConceptAttention Helbling et al. (2025) which focus on concept localization and interpretability, **MIRAGE** uses model-internal representations directly for *reference-free generation evaluation*. Tinaz et al. Tinaz et al. (2025) study interpretable concepts in diffusion activations using sparse representations, while TIDE Huang et al. (2025b) is particularly close in its use of sparse autoencoders and temporal analysis in diffusion transformers. In contrast, MIRAGE converts sparse internal activations into two evaluation signals—*cross-seed spatial consistency* and *activation concentration*—and combines them into a single reference-free score for creative text-to-image generation.

C Experimental Setup

We evaluate **MIRAGE** on three transformer-based text-to-image models: **SD3**, **PixArt**, and **FLUX**. For SD3 and PixArt, we extract image-token representations from transformer blocks $\{4, 8, 12, 16, 20\}$, while for FLUX we consider representative early and late blocks from both its double-stream and single-stream stages. For each generation, representations are collected across the specified denoising steps and four stochastic seeds $\{42, 45, 12, 1000\}$. A separate sparse autoencoder is trained for each selected transformer block, with the block-specific SAE subsequently reused across the evaluated denoising steps. Each SAE uses a latent dimension of $D = 4096$ and is trained for 10 epochs with Adam (learning rate 10^{-3}, batch size 1024) using mean-squared reconstruction loss with an L_1 sparsity coefficient $\lambda = 10^{-3}$. The decoder columns are normalized after each optimization step. **MIRAGE** is then computed from cross-seed centroid consistency and spatial activation concentration. All experiments were conducted using **2 CPU cores and one NVIDIA H100 GPU with 160 GB memory**. Processing the full set of **4,200 training samples** required approximately **24 hours**. During SAE training and representation processing, GPU memory consumption was approximately **23–26 GB for PixArt and SD3**, respectively, and **33 GB for FLUX**. Human evaluation was conducted separately and was not used for SAE training or block selection.

D Additional Experimental Results

This section provides additional experimental analyses supporting the main results. We examine the individual components of **MIRAGE**, the choice of aggregation function, cross-architecture behavior, the human evaluation protocol and inter-annotator agreement, correlations with existing reference-free metrics, and qualitative failure and prompt-level visualizations.

Table 3: Detailed **MIRAGE** component statistics on the balanced evaluation set of 193 samples per model.

Model	$CD \uparrow$	$C \uparrow$	**MIRAGE** $\uparrow$
FLUX	0.9344	0.6792	0.7866
SD3	0.9543	0.6827	0.7959
PixArt	0.9454	0.6659	0.7814

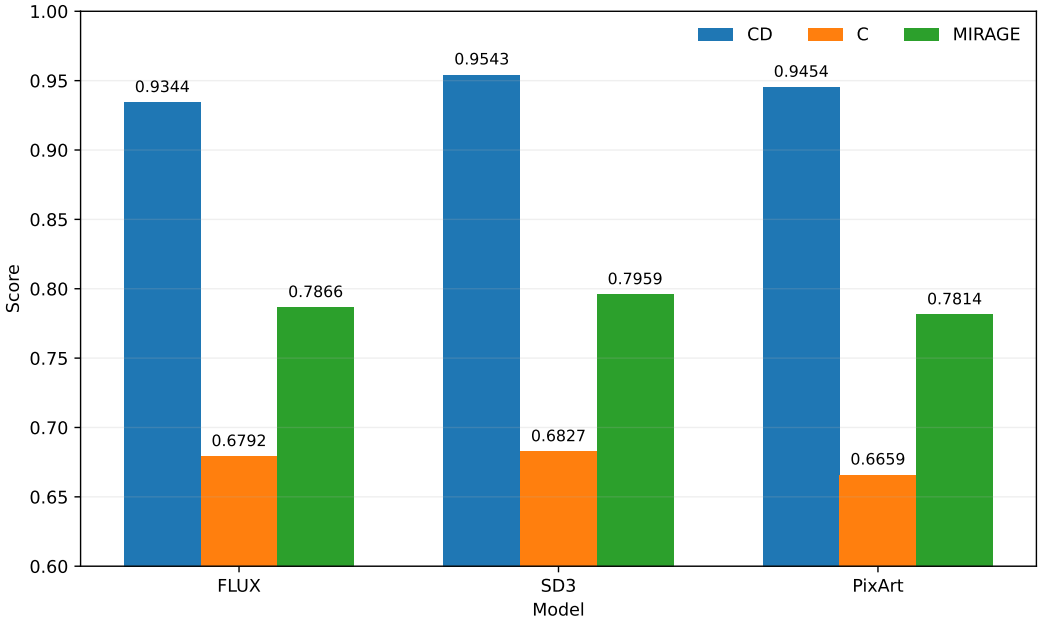

Figure 2: Component-wise analysis of MIRAGE on the balanced evaluation set. CD denotes cross-seed centroid consistency, C denotes spatial activation concentration, and MIRAGE is their equal-weight combination. The component values exhibit complementary behavior across the evaluated models.

D.1 **MIRAGE** Component Analysis

MIRAGE combines two complementary internal signals: cross-seed centroid consistency (CD) and activation concentration (C). Table 3 reports their detailed values for the three evaluated diffusion-transformer architectures.

The component values demonstrate that the two signals capture different aspects of the internal representation. In particular, cross-seed consistency characterizes the stability of the spatial organization of the dominant sparse activation under stochastic variation, whereas activation concentration characterizes its spatial localization within an individual generation.

This distinction is also visible at the block level. For FLUX, the early double-stream blocks achieve the highest centroid consistency ($CD = 0.9539$), whereas the late single-stream blocks achieve the highest activation concentration ($C = 0.7088$). Thus, high spatial stability does not necessarily imply high spatial concentration, supporting the use of both signals in the final metric.

D.2 Aggregation Ablation

The final **MIRAGE** score combines two complementary internal signals, CD and C, which capture cross-seed spatial consistency and activation concentration, respectively. Since both properties are necessary for a strong internal spatial signal, the aggregation function should penalize cases in which one component is substantially weaker than the other.

Table 4: Ablation of the **MIRAGE** aggregation function. All aggregation rules use the same CD and C values. The harmonic mean is used in the final formulation because it penalizes imbalance between the two complementary components.

Model	Arithmetic	Geometric	**Harmonic**
FLUX	0.8068	0.7965	**0.7864**
SD3	0.8185	0.8071	**0.7959**
PixArt	0.8057	0.7934	**0.7813**
Mean	0.8103	0.7990	**0.7879**

Table 5: FLUX block-wise **MIRAGE** statistics on the balanced evaluation subset.

Block group	CD	C	**MIRAGE**
Double-stream early	0.9539	0.6697	0.8118
Double-stream late	0.9248	0.6712	0.7980
Single-stream early	0.9380	0.6669	0.8024
Single-stream late	0.9208	0.7088	0.8148

We therefore compare three symmetric aggregation functions: the arithmetic mean, geometric mean, and harmonic mean,

$$M_{\text{arith}} = \frac{CD + C}{2}, \tag{33}$$

$$M_{\text{geo}} = \sqrt{CD \cdot C}, \tag{34}$$

and

$$M_{\text{harm}} = \frac{2CD \cdot C}{CD + C}. \tag{35}$$

The three formulations use exactly the same CD and C values and differ only in how the two signals are aggregated. The resulting scores are shown in Table 4.

The arithmetic mean produces the largest numerical value, while the geometric and harmonic means increasingly penalize disparities between the two components. This difference is expected from the properties of the aggregation functions and is not itself evidence that one formulation provides better evaluation. We select the harmonic mean because the intended interpretation of **MIRAGE** requires both cross-seed spatial consistency and activation concentration to be strong. The harmonic formulation consequently implements this requirement directly: a high value of one component cannot completely offset a substantially lower value of the other, while equal component values are preserved without introducing an additional weighting parameter.

D.2.1 Cross-Architecture Analysis

We further examine the behavior of **MIRAGE** across transformer blocks. This analysis complements the model-level results in the main paper by showing how the two internal components vary according to the location of representation extraction.

Across all three architectures, the component values vary with the transformer location. The variation is not identical for CD and C, further supporting their complementary roles. The detailed denoising-step behavior is analyzed separately in Appendix E.

D.3 Human Evaluation Protocol and Inter-Annotator Agreement

Because human annotation is costly, we use a balanced test set of 579 prompt–model samples, comprising 193 samples for each of FLUX, SD3, and PixArt. The complete annotation pool contains

Table 6: SD3 block-wise **MIRAGE** statistics.

Block	CD	C	**MIRAGE**
Transformer 4	0.9787	0.6865	0.8326
Transformer 8	0.9629	0.6843	0.8236
Transformer 12	0.9442	0.6826	0.8134
Transformer 16	0.9456	0.6808	0.8132
Transformer 20	0.9403	0.6795	0.8099

Table 7: PixArt block-wise **MIRAGE** statistics.

Block	CD	C	**MIRAGE**
Transformer 4	0.9426	0.6736	0.8081
Transformer 8	0.9466	0.6649	0.8057
Transformer 12	0.9386	0.6612	0.7999
Transformer 16	0.9483	0.6663	0.8073
Transformer 20	0.9512	0.6635	0.8073

600 individual ratings over 579 unique prompt–model samples. Each sample is evaluated along four dimensions: Semantic Fidelity, Attribute Presence, Compositional Correctness, and Narrative Fidelity. The overall human score is computed as the mean of the four dimension-level scores. The balanced subset is used for comparisons across models so that differences in annotation coverage do not affect the metric comparisons. Inter-annotator agreement is evaluated on the subset of samples that received two independent annotations. Because the human scores are ordinal rather than categorical, we discretize the scores into ordered rating levels and compute weighted Cohen's κ, which accounts for the magnitude of disagreements.

Table 8 reports the resulting agreement values. Agreement is highest for Compositional Correctness, while the other dimensions exhibit more modest agreement. This variation is expected for subjective evaluation of open-ended creative generations and motivates reporting the individual evaluation dimensions in addition to the aggregate score.

D.4 Human Scores and Metric Correlations

Table 9 reports the mean human scores for each evaluation dimension on the balanced test set.

D.5 Failure Cases of Existing Metrics

Figure 3 presents representative cases where reference-free output-level metrics produce scores that do not align well with visible properties of the generated images. CLIPScore assigns the three examples scores within a narrow range (0.137–0.213), including a gibberish-text image receiving a higher score than a more plausible creative generation. BLIP-2 assigns a perfect score (1.000) to the illegible-text image and a very high score (0.999) to an image in which the prompt-relevant content is only weakly represented, while assigning a much lower score (0.113) to a coherent, on-theme generation. VQAScore similarly assigns a high score (0.865) to the illegible-text image and 0.774 to a visually imperfect leaf image, while the more relevant child–bird generation receives a lower score of 0.531.

In contrast, **MIRAGE** assigns consistently low scores across these examples (0.26–0.34), including the cases receiving unusually high scores from the output-level metrics. This suggests that the model-internal spatial signals captured by **MIRAGE** are less affected by the salient visual or textual cues that can inflate individual output-level scores. These qualitative examples motivate using model-internal spatial information as a complementary signal for reference-free evaluation of creative text-to-image generation.

Table 8: Weighted Cohen's κ for the doubly annotated samples.

Dimension	Weighted Cohen's κ
Semantic Fidelity	0.50
Attribute Presence	0.72
Compositional Correctness	0.60
Narrative Fidelity	0.55
Overall	0.55

Table 9: Mean human evaluation scores on the balanced test set.

Model	Semantic	Attribute	Composition	Narrative	Overall
FLUX	0.716	0.711	0.711	0.731	0.717
SD3	0.718	0.726	0.716	0.726	0.721
PixArt	0.742	0.730	0.741	0.747	0.740

D.6 PROMPT-LEVEL **MIRAGE** VISUALIZATION

To provide a complete view of how **MIRAGE** is computed for an individual generation, Figure 4 presents a representative prompt together with its generated image and the corresponding intermediate activation maps. The heatmaps show the spatial activation of the selected sparse feature at the evaluated representation locations, while the centroid and concentration statistics summarize the spatial organization of these activations.

The visualization connects the final **MIRAGE** score to the underlying model representations: activation concentration describes the spatial localization of the activation within each generation, whereas cross-seed centroid consistency measures the stability of its spatial organization across stochastic generations. This provides a direct visual interpretation of the quantities used in the final score.

To qualitatively illustrate the model-internal signals used by **MIRAGE**, Figure 5 and Figure 6 show stochastic generations and the corresponding SAE-derived spatial activation maps across transformer blocks.

E TEMPORAL REPRESENTATION DYNAMICS

Because the same block-specific SAE is reused across the evaluated denoising steps, the spatial statistics obtained at different steps can be compared within a common SAE latent coordinate system. We therefore analyze the evolution of the centroid-consistency signal $CD^{(l,t)}$ across the denoising trajectory.

For each transformer block, we obtain a trajectory

$$\mathcal{D}^{(l)} = \left\{ CD^{(l,t)} \right\}_{t \in \mathcal{T}}. \tag{36}$$

This temporal analysis is used to characterize how the spatial organization of the dominant sparse feature changes during denoising. It is not included as an additional term in the final **MIRAGE** score.

Table 10: Pearson correlation between MIRAGE and individual human evaluation dimensions.

Model	Semantic	Attribute	Composition	Narrative
All	0.112	0.087	0.122	**0.166**
FLUX	0.083	0.078	0.121	0.089
SD3	0.124	0.087	0.167	**0.218**
PixArt	0.148	0.111	0.077	**0.208**



Figure 3: Representative failure cases of existing reference-free metrics. The examples illustrate cases where output-level scores do not fully reflect prompt-specific properties visible in the generated images. **MIRAGE** provides an additional model-internal perspective based on the spatial organization of intermediate representations.

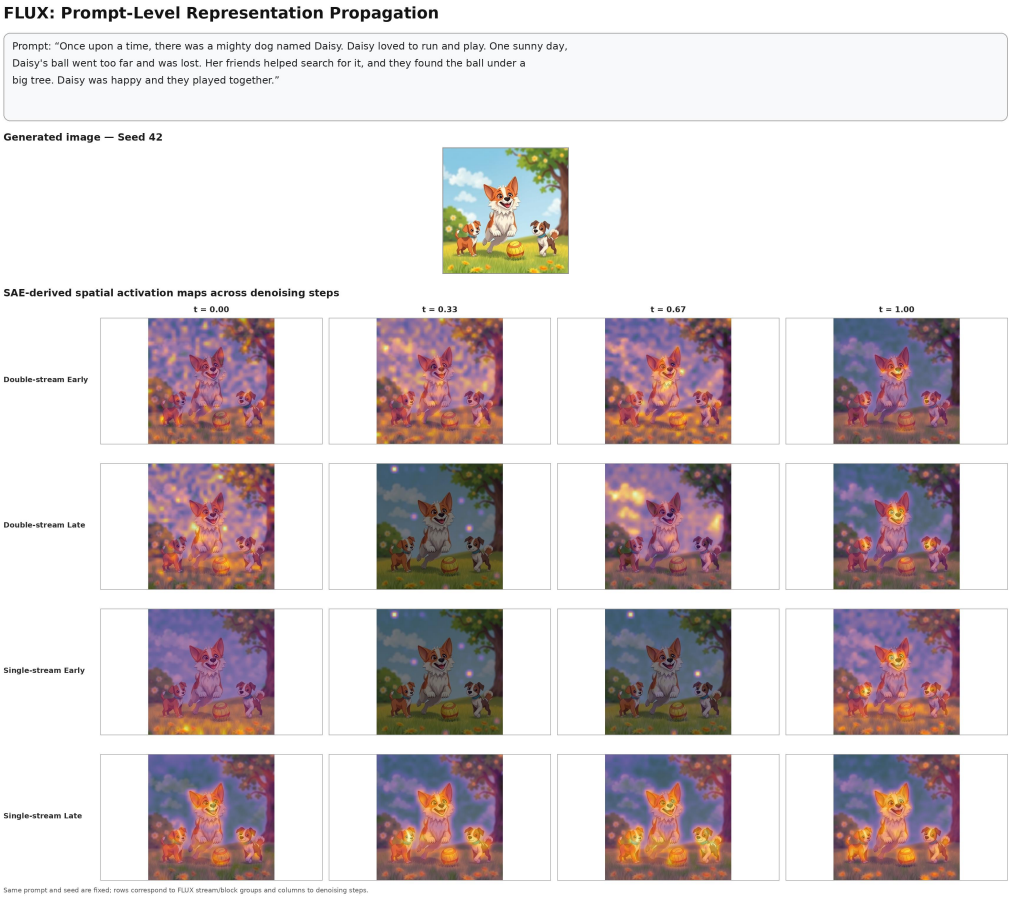

Figure 4: Prompt-level visualization of **MIRAGE**. The figure shows the input creative prompt, generated image, intermediate sparse-feature activation heatmaps, spatial centroids, component statistics, and final **MIRAGE** score. The visualization illustrates how the model-internal spatial signals used by **MIRAGE** are derived and combined for an individual generation.

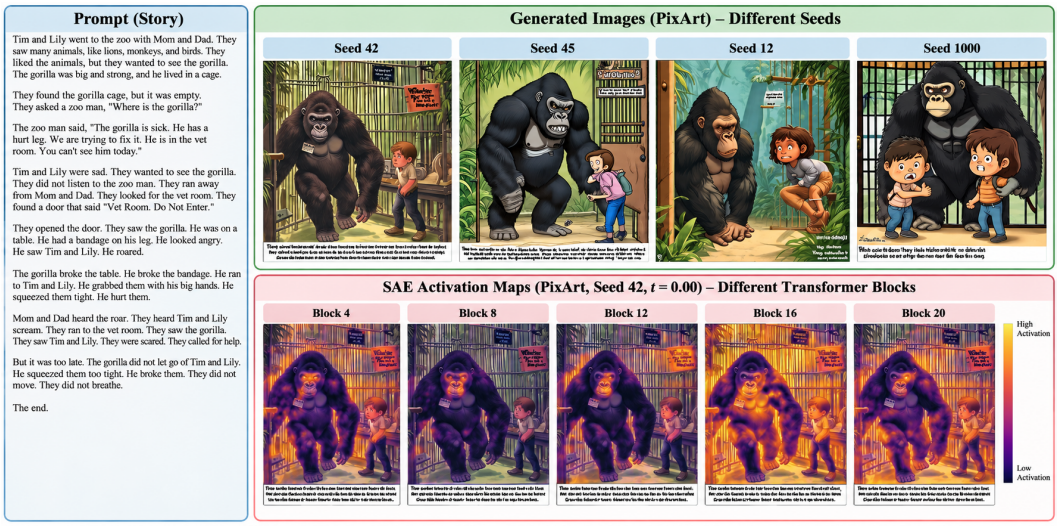

Figure 5: Qualitative visualization for PixArt showing four stochastic generations of the same prompt and SAE-derived spatial activation maps across transformer blocks at $t = 0.00$.

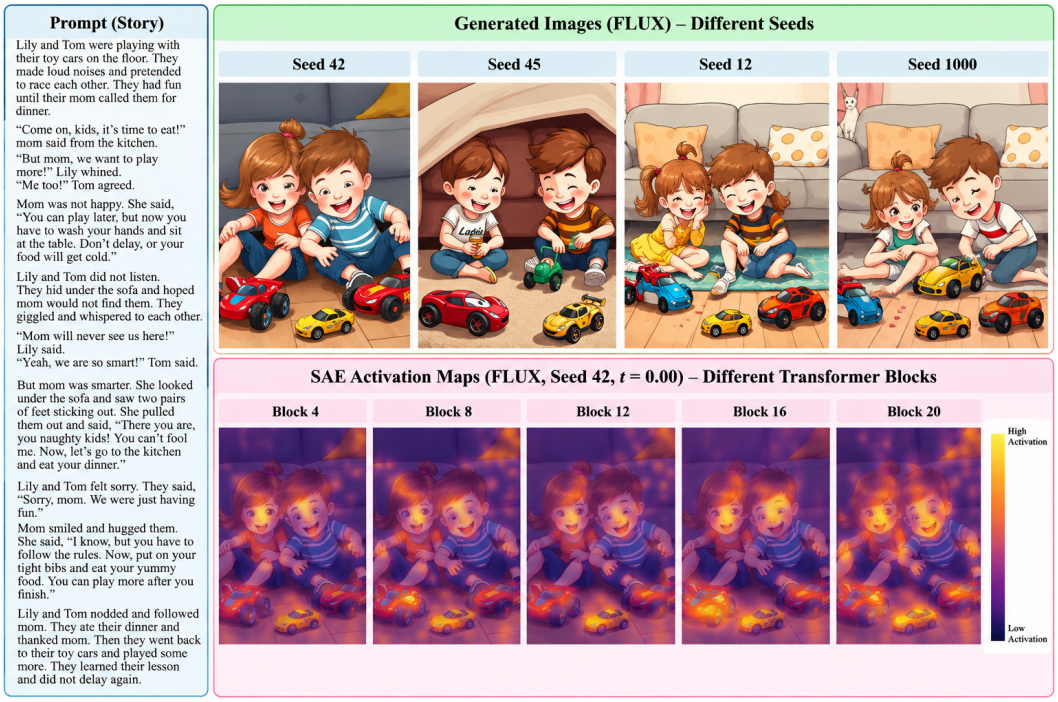

Figure 6: Qualitative visualization for FLUX showing four stochastic generations of the same prompt and SAE-derived spatial activation maps across transformer blocks at $t = 0.00$.