Evaluation for Text-to-Image Generation from a Creativity Perspective

Anonymous ACL submission

Abstract

In recent years, driven by advancements in diffusion process, Text-to-Image (T2I) models have rapidly developed. However, evaluating T2I models remains a significant challenge. While previous research has thoroughly assessed the quality of generated images and image-text alignment, there has been little study on the creativity of these models. In this work, we define the creativity of T2I models based on previous definitions of machine creativity. We also propose corresponding metrics and design a method to test the reliability of the metric. Additionally, we create a fully automated pipeline that, through text vector retrieval and the text synthesis capabilities of large language models (LLMs), can convert existing image-text datasets into benchmarks needed for evaluating creativity. Finally, we conduct a series of tests and analyses on the evaluation methods for creativity and the factors influencing the creativity of the models. The code and benchmark will be released.

1 Introduction

Inspired by diffusion process, researchers have designed a series of Text-to-Image (T2I) models based on this theory, which exhibit outstanding performance and have significantly contributed to the development of image generation, such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024), FLUX (Labs, 2024) and DALL-E3 (Betker et al., 2023), demonstrating powerful capabilities in generating relevant visual images from textual input. Despite the rapid advancement of image generation, a significant challenge remains: automated image evaluation (Lin et al., 2025; Tu et al., 2024), where the primary focus is typically on image quality and text-image consistency.

In image quality evaluation, Inception Score (Salimans et al., 2016) measures diversity with a pre-trained Inception network, while FID (Heusel et al., 2017) compares the distribution of generated and real images. For text-image consistency, approaches typically involve comparing generated captions with human-annotated ones (Hong et al., 2018), or utilizing the CLIP Score (Brooks et al., 2023; Li et al., 2024; Wu et al., 2023; Esser et al., 2024) adopted CLIP (Radford et al., 2021), which quantifies the cosine similarity between image and text embeddings.

T2I models are capable of generating high-quality, stylistically distinct images, achieving high scores on existing evaluation metrics, however, the evaluation perspectives discussed above give limited attention to the creativity of the model. Evaluating creativity is crucial for measuring a model's ability to generate interesting content. This is especially important in assisting professionals in fields such as art, design, and innovation. At the same time, it extends the practical value of the model, enabling it to contribute to the development of industries such as advertising, fashion, and entertainment. Karampiperis et al. (2014) demonstrated that the creativity exhibited in text artifacts can be predicted using appropriate formulations of computational creativity metrics. Aghazadeh and Kovashka (2024) defined the creativity of images as their uniqueness in advertisement image generation and exhibited that current T2I models face challenges when it comes to generating creative outputs. Building upon the broader definitions of machine creativity (Franceschelli and Musolesi, 2024) in previous works, we extend this concept to T2I models, providing a specific definition for them, which is divided into three components: **Value**, **Novelty**, and **Surprise**. **Value** refers to whether the images align to human's instruction. **Novelty** refers to the uniqueness of the image in relation to other images generated by the same model. **Surprise** refers to whether the images contain unexpected or surprising content.

Based on the definitions we proposed, we estab-

lished corresponding metrics, benchmarks, and a pipeline capable of automatically generating benchmarks based on existing image-text datasets, which create a benchmark where one prompt corresponds to multiple images by clustering and merging similar texts from text-image pairs. Through multiple experiments, we tested the proposed metrics and demonstrated their feasibility. Additionally, we explored various factors that influence the evaluation of model creativity. On the generated benchmark, we tested the creativity of different versions of Stable Diffusion and observed that while Value consistently increased with each version, surprisingly, both Novelty and Surprise did not follow the same upward trend and, in fact, showed a decline. This finding underscores the importance of evaluating model creativity.

In summary, the key contributions of our study are threefold:

1. Based on the general concept of machine creativity, we define the creativity of T2I models as consisting of Value, Novelty, and Surprise, and have designed evaluation methods along with relevant metrics.

2. We have designed a fully automated pipeline that can convert existing image-text datasets into the benchmark required for evaluating creativity, without the need for manual intervention.

3. We tested our proposed metrics and demonstrated their feasibility. Furthermore, we evaluated different T2I models on the generated benchmark and found that Novelty and Surprise did not increase with version updates; instead, they decreased. This also highlights the importance of assessing creativity.

2 Related Works

2.1 T2I Models

The development of deep learning has made the transformation from text to image possible, and the advancement of T2I models has been rapid. Reed et al. (2016) was the first to introduce Generative Adversarial Networks (GANs) to the text-to-image task. Subsequently, numerous works (Zhu et al., 2019; Park et al., 2019; Kang et al., 2023; Sauer et al., 2023) have been based on GANs, continuously optimizing the performance of T2I models for this task. However, T2I models based on diffusion process soon gained widespread attention, leading to the emergence of numerous impressive models. Rombach et al. (2022) presented a latent diffusion model, which significantly improved training efficiency and has the capability to generate high-quality, high-resolution images. Compared to previous versions of Stable Diffusion, Stable Diffusion XL (Podell et al., 2023) designs a model with more parameters and introduces a refinement model to improve details. The model has achieved significant performance improvements over past models. Stable Diffusion 3 (Esser et al., 2024) improves existing noise sampling techniques and introduces a new transformer-based (Vaswani, 2017) model architecture, resulting in further performance enhancements.

2.2 T2I Metrics & Benchmarks

In recent years, designing automatic evaluation metrics to assess the quality of machine-generated images has always been a topic of great interest among researchers in the field of computer vision. Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (Heusel et al., 2017) are the most widely adopted image quality metrics. The former extracts visual features from generated images using a pre-trained Inception-V3 model (Szegedy et al., 2016) to evaluate image diversity. The latter compares these extracted features with those of "gold" images to assess image fidelity. CLIPScore (Hessel et al., 2021) is based on computing the cosine similarity between image and text embeddings, as a metric for image-text alignment. VQAScore (Lin et al., 2025) evaluates the alignment between an image and a text prompt by leveraging the latent knowledge of large models. It calculates the probability that the model answers “Yes" to the question “*Does this figure show ‘text’?*".

Additionally, a high-quality benchmark is urgently needed for evaluating T2I (Text-to-Image) models. Visual Genome (Krishna et al., 2017) and MSCOCO (Lin et al., 2014) are widely used datasets for computer vision research, consisting of large-scale real-world scenes annotated for tasks such as object detection, captioning and also for evaluating image quality and image-text consistency. TIFA v1.0 (Hu et al., 2023) is a benchmark that includes 4k diverse text inputs and 25k questions across 12 categories for T2I faithfulness evaluation. DSG-1k (Cho et al., 2023) encompasses a broad spectrum of fine-grained semantic categories,

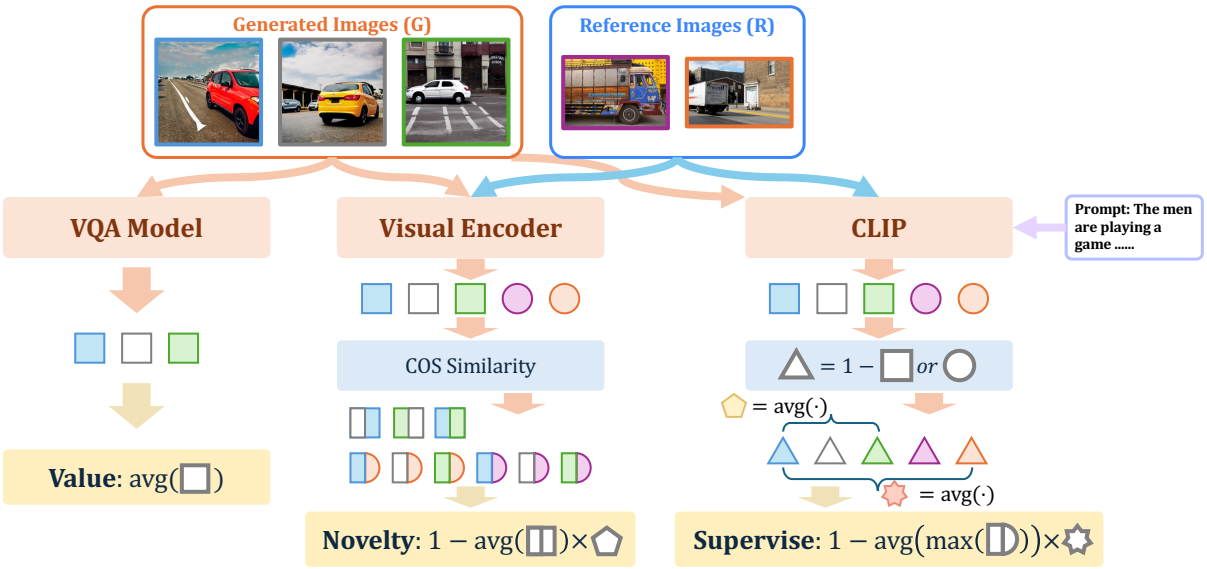

Figure 1: An illustration of metric calculation process, including **Value**, **Novelty** and **Surprise**. Firstly, We encode the images with Visual encoder and compute the cosine similarity between the vectors of the generated images, and also compute the cosine similarity between the vectors of the generated images and the reference images. Simultaneously, we calculate the text-image similarity by CLIP, interpreting this similarity as the proportion of the semantic content in the prompt relative to the overall visual semantic content. This allows us to estimate the proportion of the visual semantics that lies outside the scope of the prompt. By using a weighted approach, we compute a more reasonable distance between the images to measure Novelty and Surprise. Additionally, we calculate the mean of the VQAScore as Value.

ensuring a balanced distribution throughout.

3 Creativity Evaluation

3.1 Creativity Definition for T2I Model

Franceschelli and Musolesi (2024) considered Boden's criteria for studying machine creativity defined as "the ability to come up with ideas or artifacts that are new, surprising and valuable" (Boden, 2004). **Value** encompasses utility, performance, and attractiveness, and is connected to both the quality of production and its societal acceptance (Maher, 2010). **Novelty** refers to the degree of difference between the created artifact and others within its class (Ritchie, 2007). **Surprise** refers to how much a stimulus deviates from expectations (Berlyne, 1973).

Based on the widely accepted definition of creativity in previous research, we have provided a specific definition for Text-to-Image models. **Value** refers to whether the generated images include the content mentioned in the prompt. **Novelty** refers to the unique aspects of a image compared to other images generated by the same model. **Surprise** refers to whether the generated images contain new content that exceeds common human knowledge.

3.2 Creativity Metric

3.2.1 Value

To evaluate whether the images generated by a model effectively capture the content described in the prompts, we chose to use VQAScore as the evaluation metric. Compared to the commonly used CLIP model, CLIP is trained via contrastive learning to establish a one-to-one correspondence between images and text. In contrast, VQAScore evaluates the likelihood of a "Yes" response from a Large Vision Language Model (LVLM) when queried with relevant questions. LVLMs are typically trained on large-scale datasets and support more flexible question forms, whereas CLIP is limited to calculating relatively rigid image-text similarity. We take the average VQAScore of a set of generated images as the score of Value for the model when generating this set of images, the formula as follows.

$$Value = \frac{1}{N}\sum_{n=1}^{N}\text{VQAScore}(i_n^g, t) \quad (1)$$

where i_n^g represents the n^{th} generated image, while t denotes the prompt for image generation, and N is the number of generated images.

3.2.2 Novelty

According to the definition, we aim to evaluate whether there are significant differences between images generated multiple times by the same model under the same prompt. We measure the visual semantic distance between generated images with visual encoder, which serves as the basis for calculating Novelty. As shown in Fig. 1, we also calculate the average of the image-text similarity between the generated images and the prompt, approximating this as the proportion of the prompt's semantics represented within the visual semantics. This allows us to derive the proportion of other semantics excluding the prompt in the visual content. Since all the generated images include the content of the prompts, our evaluation focuses on assessing the content beyond the prompts, which is our primary focus of interest. Specifically, we aim to evaluate the semantic distance of non-prompt content generated across a T2I model's multiple attempts for generation. By leveraging the semantic proportion, we approximate the similarity of the content outside the prompts. Finally, the average semantic distance of the content out of prompt is calculated as Novelty score by averaging the similarity scores.

$$d_n^g = Encoder(i_n^g) \tag{2}$$

$$Prop_{nov} = 1 - \frac{1}{N}\sum_{n=1}^{N} CLIP(i_n^g, t) \tag{3}$$

$$Novelty = 1 - \frac{2}{N^2 - N} * \sum_{n=1}^{N}\sum_{j=n+1}^{N} cos\_sim(d_n^g, d_j^g) * Prop_{nov} \tag{4}$$

where d_n^g represents the visual embedding of the n^{th} generated image, and $Prop_{nov}$ in Novelty denotes the estimated proportion of similarity for content outside the prompt.

3.2.3 Surprise

Similar to Novelty, we aim to evaluate whether the images generated multiple times by the model under the same prompt can contain content that exceeds human expectations. The Surprise evaluation process is similar to Novelty, with two main differences. One difference is that we introduce a reference image set. As mentioned in section 3.1, the Surprise metric is designed to evaluate whether the content of an image generated by a T2I model is beyond common knowledge. The reference image set consists of real images that not only contain the prompt's content but also include common content associated with the prompt. The Surprise is calculated by measuring the distance between the generated images and these reference images. Similarly to Novelty, we aim to evaluate the distance between the content beyond the prompt, as the prompt content is mandatory for all images. Therefore, we also introduce CLIP. The other difference is that, unlike Novelty, maximum pooling substitute average pooling when calculating the similarity between a generated image and multiple reference images. This is because our expectation for Surprise is more stringent; once the content is predictable, it is no longer a Surprise.

$$Prop_{surp} = 1 - \frac{1}{N+S}[\sum_{n=1}^{N} CLIP(i_n^g, t) + \sum_{s=1}^{S} CLIP(i_s^r, t)] \tag{5}$$

$$Surprise = 1 - \frac{1}{N} * \sum_{n=1}^{N} \max_{s \in S} cos\_sim(d_n^g, d_s^r) * Prop_{surp} \tag{6}$$

where i_s^r and d_s^r represent the s^{th} reference image and its visual embedding respectively, and S is the number of reference images.

3.3 Benchmark & Generation Pipeline

To evaluate the model's creativity, we have constructed a fully automated process that can transform existing image-text datasets into benchmarks required for assessing creativity, as depicted in the Fig 2.

By encoding the text in the image-text pairs of the dataset and then clustering them, all text vectors are divided into n categories, where n depends on the desired size of the benchmark. Next, the pipeline randomly select one text from each cluster and, based on similarity calculations, find the $k-1$ most similar prompts within the same cluster. The value of k depends on the number of reference images needed for evaluating Surprise. Then retrieve the images corresponding to these k prompts to serve as reference images. Finally, the pipeline merge the k prompts into a single prompt with a LLM, ensuring that the merged prompt corresponds

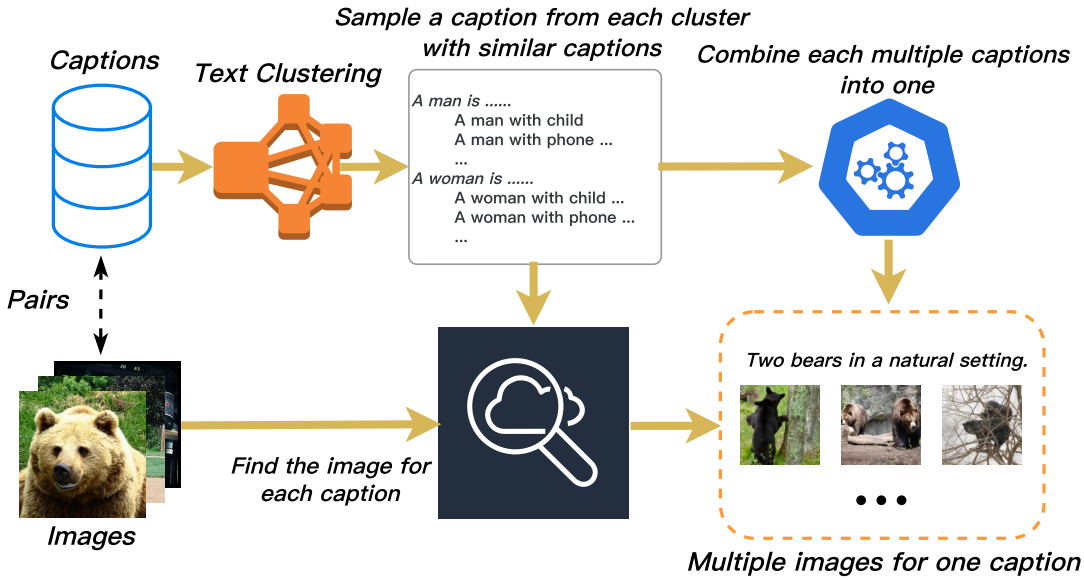

Figure 2: An illustration of fully automated benchmark generation pipeline. First, convert the text in the text-image pairs into vectors. Then, cluster all the texts into multiple categories. From each category, randomly select one text and find the most similar texts within the same category. Next, find the images corresponding to all these texts. Finally, to ensure that the prompt can correspond to all images, use a large language model to merge the similar texts into one. This process creates a benchmark where one prompt corresponds to multiple images.

to all the reference images, with the prompt, "Here are some captions. '*{captions}*' Please find what these captions have in common, don't have to describe the difference between them, DO NOT use generalisations such as various, different and so on and write it in one caption. Please only answer the caption without anything else.". In this paper, the value of k is 6, resulting in a benchmark consisting of 384 prompts and their reference images based on MSCOCO (Lin et al., 2014).

4 Experiments

4.1 Test for Metric

Through extensive experiments and consistency tests with human judgments, Fu et al. (2024) found that the DINO model is capable of capturing subtle differences in visual semantics. Therefore, we choose the DINOv2 large model (Oquab et al., 2023) as the visual encoder when the evaluating the metric. Although DINO has been experimentally proven to capture visual semantics (Fu et al., 2024), we conducted further tests on our DINO-based metrics. For the Value metric, we directly use the VQAScore, so no additional testing is required. In our subsequent test experiments, we used the FLUX API provided by Alibaba to generate high-quality images for testing.

The essence of the evaluation process for Novelty and Surprise is fundamentally consistent, with the main difference that Novelty involves comparing generated images with each other, while Surprise involves comparing them with references. We designed a method, illustrated in Fig. 3, to test whether the Novelty metric can distinguish between image sets with different levels of Novelty. For evaluation, we set the T2I model to run six times to generate six different images. We predefine four levels of Novelty image sets, ranging from low to high, using an original prompt, two enriched prompts, three enriched prompts, and six enriched prompts, respectively. We enrich a prompt through LLM while retaining its original semantics. By altering the prompt, we force the T2I model to generate images containing the original prompt content in different scenarios. If we generate six images with an original prompt, these images will be quite similar. However, if the model generates with six enriched prompts, each generating one image, these six images will be significantly different. With two enriched prompts, each generating three images, results in six images with two groups, where the images between the groups are more different and in the same group are more similar. Similarly, using three enriched prompts follows the same logic.

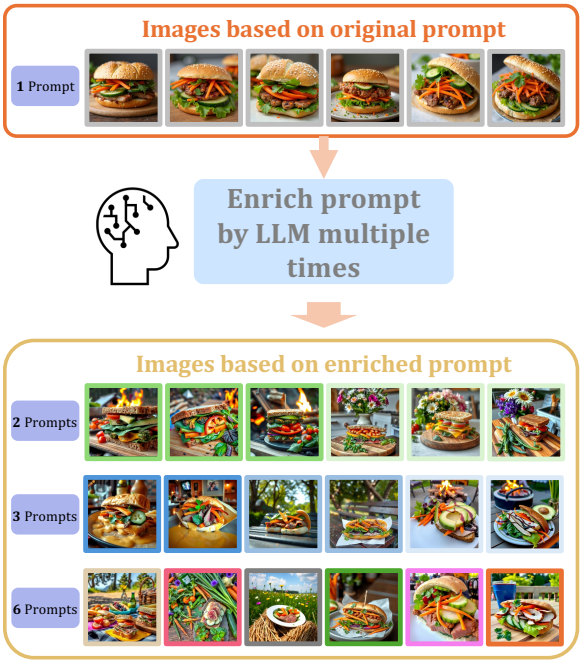

Figure 3: An illustration of the method for testing metric. Enhancing the content of original prompts through LLM while preserving their original semantics, aiming to enable the model to generate content that is richer compared to the original prompts. This approach is designed to simulate the outputs of models with varying levels of Novelty by controlling the number of enriched prompts.

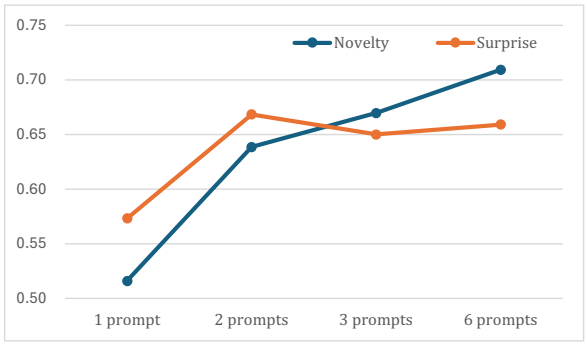

Figure 4: Test results for Novelty and Surprise. As the number of enriched prompts increases, Novelty also gradually rises. Additionally, Surprise is significantly enhanced when comparing image sets generated with enriched prompts to those generated with a single prompt. The aforementioned results align with our expectations.

To test the Surprise metric, the model generated two images with original prompt to serve as reference images. But it is not possible to preset rankings for Surprise. It's hard to control Surprise by adjusting the number of enriched prompts as we do with Novelty. Novelty involves comparing generated images with each other, where controlling the enriched prompts ensures that images generated under the same enriched prompt are similar, while images generated under different enriched prompts are significantly different. However, Surprise involves comparing the generated images with the reference images which are fixed. As long as the images generated from the enriched prompts are significantly different from the reference images provided by the original prompt, we can only preset this one ranking, i.e., the 2, 3, and 6 prompts image sets will rank higher than the 1 prompt image set. However, we cannot preset the rankings among the 2, 3, and 6 prompts image sets.

As shown in Fig. 4, the ranking of the results evaluated by the Novelty metric aligns with our predefined ranking, from low to high, one prompt, two prompts, three prompts, and six prompts, respectively. This demonstrates that our metric can distinguish the rankings of image sets with different levels of Novelty which is defined in section 3.1. As expected, the other image sets have significantly higher Surprise values compared to the 1 prompt image set, while the Surprise values among the other image sets are similar. In summary, our defined metrics can distinguish between the levels of Novelty and Surprise as defined in the previous section.

Model	Value	Novelty	Surprise
SD-v1-4	0.7858	0.5792	0.6232
SD-XL	0.8080	0.5511	0.6212
SD-v3med	0.8283	0.4981	0.6040

Table 1: Experimental results on benchmark. Value, which refers to the image-text alignment we have previously focused on, has gradually increased with model iterations. However, in the context of creativity, the newly introduced metrics of Novelty and Surprise show the opposite trend.

4.2 Implementation Details for Benchmark

We conducted the experiments on three typical T2I models: Stable Diffusion v1.4 (Rombach et al., 2022), Stable Diffusion XL base 1.0 (Podell et al., 2023), and Stable Diffusion 3 medium (Esser et al., 2024). For the visual encoder, as in the previous section, we selected the DINOv2 large model (Oquab et al., 2023). For the CLIP model, we chose to use CLIP ViT-Large Patch 14 created by OpenAI. We run the experiments on a single RTX 4090D. All models output at default resolutions.

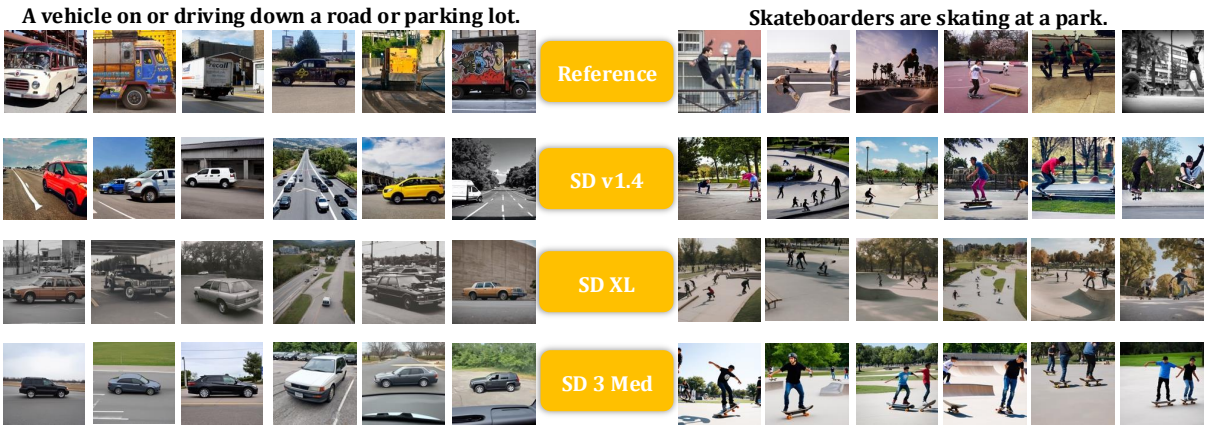

Figure 5: A generation case of benchmark. Stable Diffusion v1.4 demonstrates considerable variation in the generated images. Conversely, Stable Diffusion v3 medium exhibits minimal variation, maintaining a consistent visual angle and color palette for the car, as well as uniformity in the depiction of skateboarders. This suggests that when evaluating model performance, creativity was rarely considered before.

Specifically, the output resolutions for Stable Diffusion v1.4, XL, and 3 are 512x512, 1024x1024, and 1024x1024, respectively. For calculating VQAScore, we chose LLaVA v1.5 7B (Liu et al., 2024) as the base model. The number of inference steps is 50. Guidance scale (Ho and Salimans, 2022) is 7.5.

4.3 Results on Benchmark

As shown in Table 1, under the Value metric, the value increases with the update of stable diffusion versions. This indicates that the model is increasingly able to accurately generate content that includes the prompt, aligning with the expected model improvements. However, under the Novelty and Surprise metrics, the situation is the opposite, especially for Novelty. The decrease in Novelty (up to -0.081) means that the content generated by the model tends to become more homogeneous over multiple generations and in Surprise (up to -0.019) indicates that the content generated by the model is no longer beyond people's expectations. It is clearly observed that, from Fig. 5, in the generation tasks of these two prompts, Stable Diffusion v1.4 exhibits significant variation in color schemes, visual angles, and compositional elements across multiple generations. In stark contrast, Stable Diffusion v3 medium shows little variation, with the visual angle and color of the car remaining largely consistent, and the content related to skateboarders following the same pattern. From this perspective, it highlights the importance of evaluating creativity.

	Value	Novelty	Surprise
Baseline	0.7858	0.5792	0.6232
w/ different seeds	0.7854	0.5849	0.6271
w/ 20 images	0.7863	0.5749	0.6249
w/ guidance 12.5	0.7872	0.5645	0.6240
w/ guidance 5	0.7782	0.6025	0.6290
w/ guidance 1	0.5707	0.7801	0.7749

Table 2: Experimental results on the impact of the number of images and random seeds on the evaluation, and the effect of guidance scale on the model's creativity. It's observed that a large number of images and random seeds have little impact on evaluation. The guidance scale does influence the model's creativity; however, a very low guidance scale negatively affects Value.

4.4 Analysis

In this section, we analyze the impact of the number of generated images and different textual expressions of the same prompt semantics on the evaluation of creativity, the effect of the guidance scale on the model's creativity. We choose the Stable Diffusion v1.4 that Novelty and Surprise perform best in the benchmark experiment as the base model.

From the experimental results in Table 2, we can see that changing the random seed to generate images six times again and generating more images to evaluate the model's creativity have a negligible impact. This indicates that generating six images is sufficient to reflect the model's performance, and the performance is minimally affected by the random seed.

The default guidance scale is 7.5. Increasing

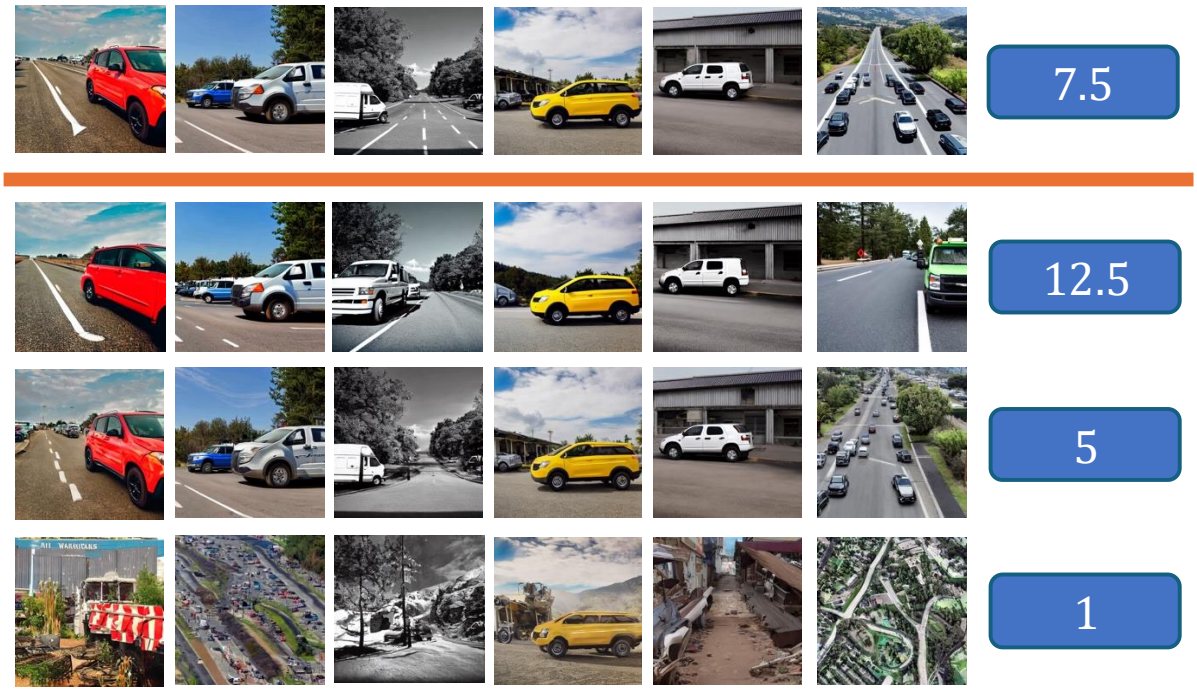

Figure 6: Cases of images generated under different guidance scale. It is evident that appropriately reducing guidance scale can enrich image content, thereby enhancing Novelty and Surprise. However, excessively lowering guidance scale, while significantly boosting Novelty and Surprise, results in images that are irrelevant to the prompt.

the guidance scale prompts the model to produce images that are more closely aligned with the text prompt. In our analysis experiments, we tested the results with scales of 12.5, 5, and 1, keeping other parameters constant. We observed that appropriately lowering the guidance scale can increases Novelty while keeping the value relatively unchanged, with a slight fluctuation in Surprise (up to -0.007 Value, +0.023 Novelty and +0.006 Surprise). However, if the guidance scale is reduced to 1, although both Novelty and Surprise increase significantly, the value drops sharply. This indicates that the high Novelty and Surprise are due to the image content deviating too much from the prompt, as shown in Fig. 6.

We sampled 50 prompts from the benchmark and used an LLM to rewrite the prompts in each group into different expressions without changing the semantics, with the prompt, "Here is a caption. '*{caption}*'. Please rewrite this caption without changing the meaning of the sentence and only answer the rewritten caption directly without anything else.". Each prompt was rewritten twice, resulting in a total of three versions including the original prompt. Each prompt generated two images, totaling six images. From the Table 3, we can find that the expression of the prompt has a minimal impact on evaluating the model's creativity under the same semantics. This result also indicates that simply altering the form of the prompt is not a feasible approach to enhancing creativity.

	Value	Novelty	Surprise
Baseline	0.7665	0.5967	0.6503
w/ rewrite	0.7684	0.6023	0.6399

Table 3: Experimental results on the effect of prompt expression on evaluation. The prompts were rewrote by LLM without altering their semantics, and the results remained largely consistent, indicating minimal impact of prompt expression.

5 Conclusion

In this paper, we explore the definition of creativity and its application in T2I models. For evaluation, we propose creativity metrics, consisting of Value, Novelty and Surprise, and an fully automatic benchmark generation pipeline. Experimental results across the generated benchmark validate creativity is a new, valuable perspective for T2I model evaluation. Furthermore, we conducted detailed analysis experiments on the influences of hyper-parameters on the evaluation of creativity.

Limitations

Despite the contributions of this work, there are several limitations that should be acknowledged. The limitations define the boundaries of our current work and suggest directions for future research.

1. When assessing the impact of the same set of images with identical semantics on the evaluation of Novelty and Surprise, we employed CLIP to approximate the semantic proportion and evaluate the distance between other semantics in different images in the set, excluding those with identical semantics. However, this method is not entirely appropriate, and a more precise approach is needed to measure the semantics we intend to compare.

2. This work focuses on evaluating the creativity of the model. For assessing the creativity of a single image, current methods may not be entirely suitable. A larger and more diverse image dataset might be necessary to support image creativity evaluation. Additionally, creative elements such as metaphors embedded within a single image may require deep exploration by large language models to be better evaluated.

Ethical Considerations

Our benchmark is derived from MSCOCO, which is licensed under the Creative Commons Attribution 4.0 License. Dinov2 large is distributed under the Apache License 2.0, while CLIP ViT-Large Patch 14 adheres to the MIT License. LLaVA 1.5 is governed by the LLAMA 2 Community License.

Our usage of these models and benchmarks in this study is strictly for academic purposes and follows license.

References

Aysan Aghazadeh and Adriana Kovashka. 2024. Cap: Evaluation of persuasive and creative image generation. *arXiv preprint arXiv:2412.10426*.

Daniel E Berlyne. 1973. Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism*, 31(4).

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*.

Giorgio Franceschelli and Mirco Musolesi. 2024. Creativity and machine learning: A survey. *ACM Computing Surveys*, 56(11):1–41.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2024. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134.

Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512. IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Dongxu Li, Junnan Li, and Steven Hoi. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Mary Lou Maher. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, pages 22–28.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Gaugan: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH 2019 Real-Time Live!*, pages 1–1.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

Graeme Ritchie. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17:67–99.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pages 30105–30118. PMLR.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. 2024. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. *arXiv preprint arXiv:2411.15488*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810.