

VLEU: a Method for Automatic Evaluation for Generalizability of Text-to-Image Models

Jingtao Cao¹, Zheng Zhang², Hongru Wang¹, Kam-Fai Wong¹

¹The Chinese University of Hong Kong

²The Hong Kong University of Science and Technology (Guangzhou)

jcao@se.cuhk.edu.hk, kfwang@se.cuhk.edu.hk

Abstract

Progress in Text-to-Image (T2I) models has significantly advanced the generation of images from textual descriptions. However, there is a lack of metrics to evaluate the models' ability to handle a diverse and extensive range of textual prompts—a key aspect of generalizability. To fill this gap, we propose the VLEU (Visual Language Evaluation Understudy) metric. VLEU leverages the power of the Large Language Model (LLM) to sample from the visual text domain, which we define as the entire set of potential inputs for the T2I task, to generate a wide variety of visual text. The images generated by T2I models from these prompts are then assessed for their alignment with the input text using the CLIP model. VLEU quantitatively measures a model's generalizability by computing the Kullback-Leibler (KL) divergence between the visual text marginal distribution and the conditional distribution over the images generated by the model. This measure serves as a quantitative metric for comparing the generalizability of T2I models, and also provides a valuable metric for assessing improvements during the finetuning process. Our experimental results demonstrate VLEU's effectiveness in evaluating the generalizability of various T2I models, positioning it as an essential metric for future research and development in image synthesis from text prompts. Our code and data will be publicly available at <https://github.com>.

1 Introduction

The emergence of latent diffusion models (LDMs) (Rombach et al., 2022) marked a significant advancement in generative models, addressing a crucial limitation that was prevalent during the era dominated by Generative Adversarial Networks (GANs) (Goodfellow et al., 2020). Unlike GANs, which were often constrained by limited expressive and computational capabilities and focused on specific tasks or datasets, LDMs, trained on extensive datasets like LAION-5B (Schuhmann et al.,

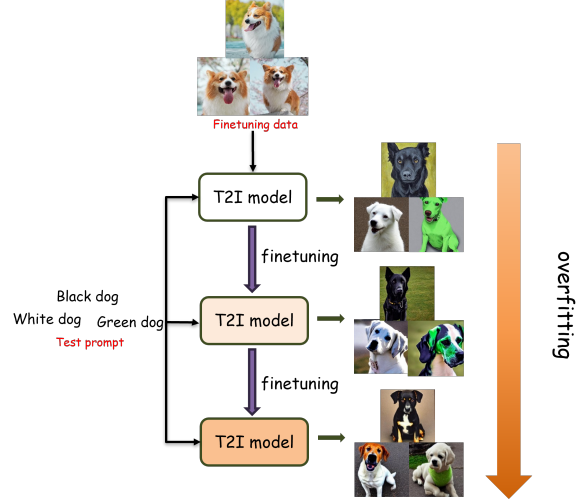


Figure 1: **The loss of generalization of a T2I model.** When fine-tuning a T2I model with images of a brown and white dog, as the fine-tuning process advances, prompts for dogs of various colors start to yield outputs that increasingly reflect the characteristics of the dog present in the training dataset. This results in generated images that deviate from the original textual description, indicating a clear case of **overfitting** and a **loss of generalization**.

2022), introduced an enhanced capacity for conditional generation across diverse scenarios. This pivotal development laid the groundwork for major strides in the field of text-to-image (T2I) generation. Notable examples include Stable Diffusion (Rombach et al., 2022), SDXL (Podell et al., 2024), Imagen (Saharia et al., 2022) and DALL-E 3 (OpenAI, 2023a), all of which have demonstrated impressive capabilities in generating detailed and contextually relevant images from textual descriptions.

When assessing the performance of T2I models, a variety of metrics are employed, including Inception Score (IS) (Salimans et al., 2016), Fréchet Inception Distance (FID) (Heusel et al., 2017), which primarily gauge the quality and diversity of the generated images. In contrast, metrics like CLIP score (Radford et al., 2021; Hessel et al., 2021), DINO

(Caron et al., 2021), and ImageReward (Xu et al., 2023) are designed to measure the semantic alignment between the generated images and the input text. Despite the effectiveness of these metrics in their respective domains, they do not fully capture a model’s generalizability, which refers to its ability to produce accurate and diverse images across a wide range of textual prompts. This aspect of T2I model performance is often evaluated through subjective human judgment, highlighting the need for a standardized measure of generalizability.

Such issue of evaluating generalizability also extends to the methodologies employed for finetuning T2I models. Techniques such as text inversion (Gal et al., 2023), LoRA (Hu et al., 2022), DreamBooth (Ruiz et al., 2023), and HiFi Tuner (Wang et al., 2023) have been instrumental in adapting pre-trained T2I models to specific subject or style. However, during the evaluation phase of these finetuned models, there is a lack of robust metrics to effectively measure the loss of generalization, which is shown in Figure 1. Many studies, including DreamBooth among others, have encountered this issue and presented it visually, ultimately relying on subjective human interpretation for assessment.

To bridge this gap, we introduce VLEU (Visual Language Evaluation Understudy) metric.¹ Firstly we define the input text prompt for the T2I task as visual text, and also the sets of potential inputs for the T2I task as visual text domain. This delineated definition is essential, as not all text is appropriate as input for T2I models. For example, the chat text from the dialogue system, when used as input for T2I models, is nonsensical. VLEU seeks to quantify the generalizability of T2I models by measuring the alignment between the visual text domain and the images generated by the T2I model conditioned on the visual text domain.

VLEU operates by calculating the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the distribution of the visual text domain and the distribution of the images conditionally generated by the model. This divergence serves as a metric for the alignment between the intended text prompts and the generated images. To facilitate this measurement, Large Language Models (LLMs) such as ChatGPT (OpenAI, 2022),

GPT-4 (OpenAI, 2023b), and LLaMA (Touvron et al., 2023a,b) are utilized to sample from the visual text domain. These descriptions are then paired with images produced by the T2I model, and the CLIP model (Radford et al., 2021) is used to evaluate the semantic congruence between each text-image pair.

The principal contributions of this work are as follows:

- We proposed the VLEU metric, an automatic evaluation designed to assess the generalizability of T2I models.
- We detailed the implementation of VLEU, which involves applying LLMs to sample visual text from the visual text domain and utilizing the CLIP model to evaluate the semantic alignment of generated images with the input visual text.
- We conducted comprehensive experiments to analyze the effectiveness of VLEU and the impact of different components in the evaluation pipeline, validating its effectiveness in quantifying T2I models’ generalizability.
- We presented two real-world case studies showcasing the practical utility of VLEU in evaluating T2I models, positioning it as a vital metric for T2I model development.

2 Background

Text-to-Image Generation: The field of image generation was once dominated by GANs (Goodfellow et al., 2020), which operate on a framework of competing networks, one generating images and the other evaluating them. However, at this stage, models based on GANs had limited expressive and computational capabilities, which constrained their generalizability across diverse tasks and datasets. For instance, StyleGAN (Karras et al., 2019) focused on generating high-quality face images but was not suitable for broader text-to-image generation. GAN research was often focused on specific domains or datasets such as CIFAR-10 (Krizhevsky, 2012), CelebA (Liu et al., 2015) and ImageNet (Russakovsky et al., 2014), rather than being capable of conditional image generation from unconstrained natural language inputs. While there were some promising works exploring conditional generation with GANs model (Mirza and Osindero, 2014; Casanova et al., 2021), significant challenges remained in bridging the gap between research prototypes and practical applications of text-to-image

¹The name VLEU is inspired by the BLEU metric, which revolutionized the evaluation of machine translation systems (Papineni et al., 2002). The nomenclature draws a parallel insofar as the T2I task can be conceptually likened to translating from text to image.

generation. The pivotal breakthrough came with the introduction of diffusion models, exemplified by DDPM (Ho et al., 2020) and DDIM (Song et al., 2020). Unlike GANs, diffusion models can synthesize high-fidelity images by gradually denoising random noise through reverse diffusion sampling. This approach provides stronger generative capabilities by leveraging the modeling power of deep neural networks. Furthermore, latent diffusion models like DALL-E (Ramesh et al., 2021) and DALL-E 2 (Ramesh et al., 2022) significantly reduced the computational costs of sampling high-resolution images. The release of the LAION-5B dataset (Schuhmann et al., 2022), containing billions of image-text pairs, provided immense amounts of training data to empower text-to-image generation. On top of the latent diffusion framework, the adoption of Transformer architectures (Vaswani et al., 2017) was another vital innovation. Models like Imagen (Saharia et al., 2022) and Stable Diffusion (Rombach et al., 2022) incorporated cross-attention layers that align textual prompts with generated image features. This mechanism enabled explicit conditioning on text descriptions and proved crucial for advancing text-to-image capabilities. The breakthrough in scaling up diffusion models to billions of parameters, combined with the effective technique of using Transformer architectures for text conditioning, has led to recent T2I models demonstrating impressive capabilities in synthesizing high-fidelity, controllable images from a wide range of textual descriptions.

Metrics for T2I Models: In evaluating T2I models, traditional metrics like Inception Score (IS) (Salimans et al., 2016), Fréchet Inception Distance (FID) (Heusel et al., 2017) focus on image quality. IS assesses the clarity and diversity of images using a pre-trained Inception network which has fixed classes trained on Imagenet (Deng et al., 2009), while FID measures the distance between feature vectors of real and generated images to gauge realism. Meanwhile, newer metrics such as DINO (Caron et al., 2021), CLIP similarity (Radford et al., 2021; Hessel et al., 2021), and ImageReward (Xu et al., 2023) offer a more nuanced assessment. DINO focuses on discerning the similarity between generated and actual images by emphasizing distinctive features. CLIP similarity metric examines the congruence between images and their textual descriptions. ImageReward gauges the aesthetic and creative attributes of generated images,

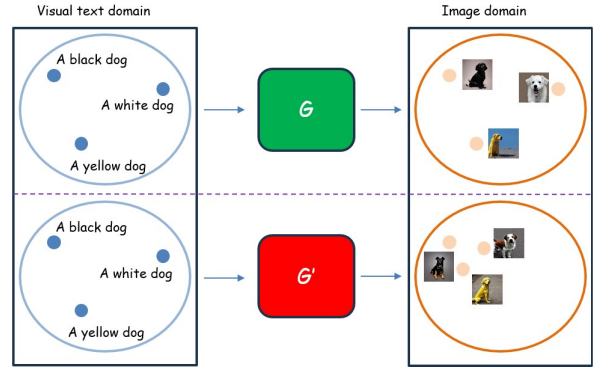


Figure 2: **Generalizability in T2I Models: A Comparative Visualization.** Model G demonstrates strong generalization, successfully producing diverse images from three distinct prompts. Conversely, G' shows limitations, overfitting to the prompt "A yellow dog" and failing to generalize to other inputs, resulting in similar outputs. Our proposed VLEU metric aims to quantify this observation.

aligning them with human aesthetic preferences. However, all these metrics primarily concentrate on either image quality or semantic alignment, revealing a gap in evaluating a model’s generalizability across varied textual prompts.

Large Language Models: The landscape of natural language processing has been revolutionized by the advent of LLMs like ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023b), and LLaMA (Touvron et al., 2023a,b). These models exhibit remarkable abilities in understanding context, generating coherent and contextually relevant text, and mimicking human conversational styles. These LLMs are not only adept in producing context-aware and coherent text but are also effectively utilized to generate prompts that guide text-to-image models. For example, recent work Visual ChatGPT (Wu et al., 2023) demonstrates using ChatGPT to automatically generate prompts for text-to-image models. The conversational nature of ChatGPT allows generating prompts with greater contextual awareness and abstraction compared to human-written prompts.

3 VLEU

How can we quantify the generalizability of a T2I model? This seems like an intractable problem at first glance. Our intuition stems from observing the phenomenon of T2I models losing generalizability during finetuning.

As shown in Figure 2, through comparison, we

Prompt Type	Example Template
Unconstrained Subject	Please imagine a random picture and describe it in one sentence.
Constrained Subject	Please imagine a picture of {class_word} and describe it in one sentence, making sure to include the word "{class_word}".
Constrained Subject with Properties	Please imagine a picture of {class_word} and describe it in one sentence, making sure to include the word "{class_word}" and words about {property}.

Table 1: **Templates used for sampling T2I prompts.** These templates serve as inputs for chat-based LLMs, such as ChatGPT, to generate T2I prompts. LLMs leverage their understanding of natural language to produce diverse and contextually relevant prompts within the visual text domain.

can find that if the T2I model can generate good images for all prompts, then the generated images should be aligned with the given prompts overall. If the model loses the ability to generate some prompts, it will cause an overall misalignment. We quantify this alignment as the KL divergence between the visual text marginal distribution and the conditional distribution over the images generated by the model, which VLEU aims to measure.

VLEU employs a three-step automatic process to measure the above KL divergence. First, text prompts are sampled from the visual text domain and used to generate corresponding images (detailed in Section 3.1). Second, the CLIP model evaluates the semantic alignment between each generated image and its original textual prompt (explained in Section 3.2). Finally, these text-image alignments are utilized for probability modeling to obtain the visual text marginal distribution and the conditional distribution over generated images. These distributions are then used to compute the final VLEU score (formulation provided in Section 3.3).

3.1 Visual Text Sampling

To initiate the evaluation process, we sample from the visual text domain - the space of potential textual inputs for the T2I task. To effectively sample from the expansive visual text domain without requiring extensive manual effort, we leverage the generalizability of LLMs to automatically generate diverse T2I prompts that closely approximate the broad sampling from the visual text domain. In this approach, we utilize LLMs to sample two types of prompts: unconstrained subject and constrained subject.

Table 1 demonstrates the templates used for sampling T2I prompts. The prompt template for constrained subjects is tailored to ensure that the generated prompts contain the same class word. This

consistency is vital when evaluating the loss of generalization in relation to a specific word. For instance, if “dog” is the class word, but the LLM replaces it with synonyms like “pooch”, “hound”, or “pup” in the T2I prompts, it could obscure the true extent to which the model’s generalizability to the word “dog” has been affected.

Additionally, despite increasing the diversity of output by adjusting parameters like the temperature in the ChatGPT API, the responses can still exhibit a degree of convergence. To counteract this and further diversify the T2I prompts, a multi-turn dialogue approach is adopted. After the initial use of the above-mentioned templates, subsequent interactions simply use the prompt “Again” to stimulate the generation of new T2I prompts. This dialogue can span up to 50 rounds, excluding the more convergent prompts typically produced in the first round.

3.2 Text-Image Scoring

Using the T2I prompts sampled from the visual text domain, corresponding images are generated by the test model under evaluation. To assess the semantic alignment between each input prompt and output image, we leverage CLIP (Contrastive Language-Image Pre-training)(Radford et al., 2021), which has demonstrated strong capabilities in matching textual descriptions to images.

Specifically, we obtain embedded representations of the text prompts and generated images from CLIP. The similarity between the text and image embeddings for each pair is then quantified using cosine similarity. Cosine similarity measures the angle between two vectors in high-dimensional space, providing a bounded similarity score that is robust to distortions from large vector magnitudes. This enables an effective assessment of how well the generated image aligns with the semantic concepts expressed in the original textual prompt.

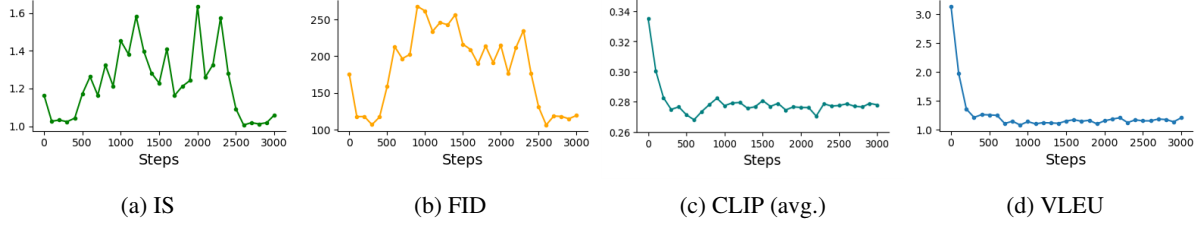


Figure 3: **Variation of different metrics during finetuning.** In this example, we finetuned SD 1.5 on 5 specific teddy bear images and sampled 25 prompts from the visual text domain covered by various teddy bears using ChatGPT 3.5. For FID, we treated the images in the dataset as the real image distribution.

3.3 VLEU Calculation

The Visual Text Sampling and Text-Image Scoring modules together provide the foundation for the VLEU metric. The VLEU metric quantitatively measures the generalizability of a T2I model by computing the divergence between two probability distributions - the marginal visual text distribution $P(x)$, and the conditional distribution $P(x|y)$ over text prompts given a generated image.

Let \mathbf{G} denote the T2I model being evaluated. Given a corpus $X = x_1, x_2, \dots, x_N$ of N textual prompts sampled from the visual text domain, the model generates corresponding image $y_i = \mathbf{G}(x_i)$. The similarity between each text-image pair (x_i, y_j) is scored using CLIP as S_{ij} .

These similarity scores are transformed into a conditional distribution over text prompts associated with each generated image via a softmax function:

$$P(x|y_i) = \text{softmax}(S_{:,i}/t) \quad (1)$$

where t is a temperature parameter. The marginal distribution is obtained by averaging over the conditionals:

$$P(x_i) = \frac{1}{N} \sum_y P(x_i|y) \quad (2)$$

Finally, the VLEU score is computed as the exponentiated expected KL divergence between the conditional and marginal distributions:

$$\text{VLEU} = \exp(\mathbb{E}_x [\text{KL}(P(x|y)|P(x))]) \quad (3)$$

Taking the exponentiation is to scale the scores into a more convenient range for comparison. This provides an interpretable measure of the model's ability to generate diverse images aligned with the visual text domain.

4 Experiments And Analysis

This section investigates the effectiveness of VLEU in evaluating T2I models and analyzes the impact of different components within the evaluation pipeline. Specifically, we conducted two experiments: (1) Analyze VLEU's effectiveness in capturing model generalizability changes during finetuning and across different T2I models, which is presented in section 4.1. (2) Assess the impact of key components like Visual Text Sampler and Text-Image Scorer on VLEU scores, which is detailed in section 4.2.

In line with our objectives, we primarily evaluated four open-source T2I models (SD 1.5, SD 2.0, SD 2.1, SDXL) under various conditions. For finetuning, we utilized the dataset provided by DreamBooth (Ruiz et al., 2023), which comprises several subsets, each containing a series of images related to a specific subject.² Besides, within all VLEU calculations, the value of temperature in Equation (1) was set to 0.01 to scale the computed results into a visually convenient range for analysis purposes.

4.1 Effectiveness Analysis

Analysis on Finetuning. We selected the finetuning process of T2I models on specific datasets, which is often considered detrimental to the models' generalizability, and tested the changes of IS (Salimans et al., 2016), FID (Heusel et al., 2017), CLIP score and VLEU during finetuning.³ Specifically, we finetuned SD 1.5 on the DreamBooth dataset and sampled 25 prompts related to each subset using GPT 3.5. For FID calculation, we treated the images in the dataset as the real image distribution. For CLIP score, we simply calculate the average clip similarity of all prompts and corresponding image.

²The dataset can be found at [google/dreambooth](https://github.com/google/dreambooth).

³For the computation of IS and FID, we utilized [toshas/torch-fidelity](https://github.com/toshas/torch-fidelity).

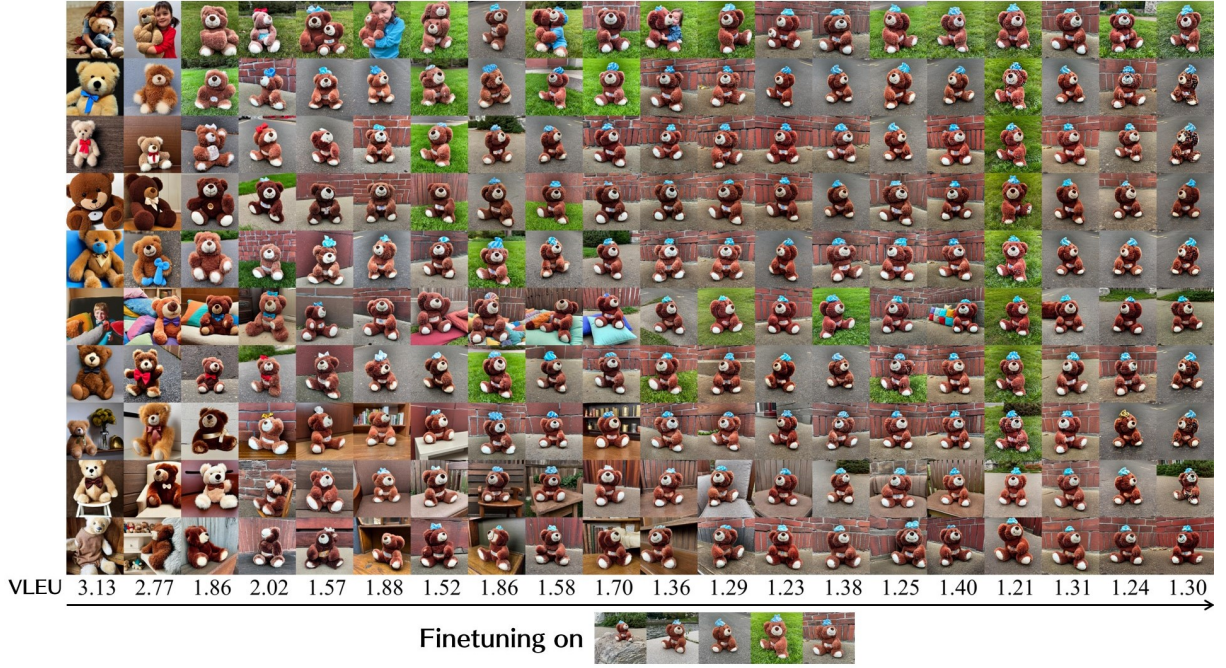


Figure 4: **Changes in VLEU of a T2I model during finetuning.** Throughout the finetuning of SD 1.5 on five particular teddy bear images, images are generated at every 20 steps using the same prompts. Meanwhile, we calculate the VLEU score at the same step. The figure indicates that the VLEU score gradually decreases as the model begins to overfit, resulting in a loss of generalization and the values align well with this trend.

Taking the finetuning on teddy bear images as an example, as shown in Figure 4, the generated images during the finetuning process tend to increasingly resemble images from the training set as the number of finetuning steps increases, indicating a decrease in model generalizability. Figure 3 illustrates the trends of IS, FID, CLIP score and VLEU throughout the finetuning process. Among these, VLEU consistently decreases, aligning more closely with the diminishing generalizability during finetuning compared to the CLIP score. However, IS and FID, which are commonly used metrics reflecting image clarity and diversity, exhibited significant fluctuations. This discrepancy can be attributed to the fact that while diversity consistently decreases during finetuning, changes in image clarity may not follow a consistent pattern.

Analysis across T2I models. We applied our VLEU metrics to comprehensively evaluate four open-source T2I models across four major visual text domains, including unconstrained, scene-focused, person-focused, and style-focused visual text. We sampled 1000 prompts in each domain using GPT 3.5. Additionally, we calculated CLIP scores, which are obtained by simply averaging CLIP similarity scores without using our VLEU calculation.

As a comparison, we also conducted a human evaluation study, where human evaluators were asked to create a variety of T2I prompts, either freely or focused on a specific subject, similar to how LLMs generate prompts. These prompts were then used to generate images with the four T2I models. The evaluators compared pairs of images generated by randomly selected pairs of models, without knowing which model produced which image, and determine which image better adhered to the prompts. We used these pairwise comparisons to compute an Elo rating (Elo, 1967) for each model, which is a method commonly used to calculate the relative levels of players in win-loss games.

As shown in Table 2, the Elo ratings derived from human evaluations were consistent with the VLEU scores, with SD 2.0 and SD 2.1 outperforming SD 1.5 and SDXL across most domains. This alignment between human evaluation and VLEU scores

demonstrates the effectiveness of our VLEU metric in assessing the generalizability of T2I models, and proves that our VLEU calculation reflects the generalizability better than simply averaging CLIP similarity scores.

For a detailed explanation of the Elo rating calculation and the human evaluation process, please

Models	Unconstrained			Scene-Focused			Person-Focused			Style-Focused		
	VLEU	Elo	CLIP	VLEU	Elo	CLIP	VLEU	Elo	CLIP	VLEU	Elo	CLIP
SD 1.5	37.18	934	0.3254	43.96	921	0.3329	48.79	894	0.3029	58.78	875	0.3355
SD 2.0	41.78	1048	0.3251	49.46	1011	0.3320	50.55	1263	0.3044	59.65	1087	0.3360
SD 2.1	42.05	1101	0.3264	50.63	1086	0.3330	50.04	980	0.3051	59.14	1111	0.3367
SDXL	39.15	917	0.3297	48.60	983	0.3352	45.68	863	0.3055	55.97	927	0.3381

Table 2: **The comparison between VLEU, CLIP, and Elo scores for different models.** We tested several common T2I models on VLEU scores across the entire visual text domain, as well as within the subdomains of scene, person, and style. We sampled 1000 prompts in each domain using ChatGPT 3.5. Additionally, we conducted a human evaluation study to compute Elo ratings for each model, based on pairwise comparisons of images generated from human-created prompts. **CLIP** refers to the average of CLIP scores obtained from text-image scoring, not using our VLEU calculation.

refer to the Appendix B.

4.2 Component Impact Analysis

In the VLEU pipeline, we focused on two key components: the Visual Text Sampler and the Text-Image Scorer. These components play pivotal roles in shaping the effectiveness and outcomes of the VLEU system. In this analysis, we investigated the impact of utilizing different models for these components on the efficacy of the VLEU pipeline.

Visual Text Sampler. We experimented with four widely used LLMs (GPT-3.5-turbo, GPT-4, LLaMA-2-7B-Chat, LLaMA-2-13B-Chat (Touvron et al., 2023b)) as the Visual Text Sampler within the VLEU pipeline and assessed their effectiveness. For each sampler, We sampled 1000 subject-unconstrained prompts and evaluated them on different T2I models. It’s worth noting that because the two LLaMA models have a weaker ability to follow instructions, they couldn’t directly output T2I prompts in the expected format under zero-shot conditions. Hence, we manually crafted initial dialogues for the first two rounds in a few-shot manner to guide the model in generating T2I prompts in the desired format. As depicted in Table 3, while different samplers exert varying influences on VLEU scores, the overall ranking of several T2I models remains largely consistent. Additionally, we observed that utilizing LLMs deemed to have better generalizability, such as GPT-4, as the Visual Text Sampler resulted in higher VLEU scores compared to those obtained with LLMs with poorer generalizability, such as LLaMA-2-7B. We hypothesize that this phenomenon stems from the weaker descriptive capability of prompts generated by LLMs with poor generalizability, leading to substantially low similarity between the generated images and these prompts. Consequently, this increases the KL divergence of each image relative to the marginal

distribution for text embeddings, thereby yielding higher VLEU scores.

Text-Image Scorer. We explored four different Text-Image Scorers (CLIP-ViT-B-16, CLIP-ViT-L-14 (Radford et al., 2021), OpenCLIP-ViT-L-14, OpenCLIP-ViT-H-14 (Cherti et al., 2023)) for computing VLEU scores. We retained 1000 subject-unconstrained prompts from GPT 3.5 along with corresponding images generated by each T2I model, only changing the Text-Image Scorer in the pipeline to compute the final VLEU scores. As shown in Table 4, it can be observed that the higher-performing scorers yield higher VLEU scores. We attribute this to the superior matching capability of the higher-performing scorers in aligning images with textual prompts. Consequently, each image exhibits greater discrepancies in scores between its own prompt and other prompts, resulting in larger KL divergences computed and thus higher final scores.

5 Case Studies Using VLEU

In this section, we present two case studies to illustrate the practical application of VLEU.

Racial Bias in T2I Models. We tested four T2I models to evaluate their VLEU scores across African, Asian, and Caucasian people. For each ethnicity, we sampled 1000 prompts using GPT 3.5⁴. The results, displayed in Table 5, indicate that all tested models achieved higher scores on Caucasians compared to Africans and Asians. This suggests that these models exhibit higher generalizability performance on Caucasians, due to significant disparities in the representation of different racial groups within the training data.

⁴The prompts are provided in the supplementary materials

T2I Model	ChatGPT 3.5	GPT-4	LLaMA-2-7B	LLaMA-2-13B
SD 1.5	37.18	36.85	58.89	47.79
SD 2.0	41.78	40.72	60.15	49.27
SD 2.1	42.05	41.48	59.12	49.41
SDXL	39.15	39.33	58.30	48.27

Table 3: **VLEU scores under different Visual Text Samplers.** For each sampler, We sampled 1000 subject-unconstrained prompts and evaluated them on different T2I models.

T2I Model	CLIP-ViT-B-16	CLIP-ViT-L-14	OpenCLIP-ViT-L-14	OpenCLIP-ViT-H-14
SD 1.5	37.18	45.86	88.82	98.83
SD 2.0	41.78	51.01	102.45	133.76
SD 2.1	42.05	52.64	104.22	132.72
SDXL	39.15	50.87	96.04	111.56

Table 4: **VLEU scores using different Text-Image Scorer.** We retained 1000 subject-unconstrained prompts from ChatGPT 3.5 along with corresponding images generated by each T2I model, only changing the Text-Image Scorer in the pipeline to compute the final VLEU scores.

T2I Model	African	Asian	Caucasian
SD 1.5	34.45	25.89	<u>89.96</u>
SD 2.0	36.59	29.50	<u>92.99</u>
SD 2.1	38.03	28.13	95.26
SDXL	32.88	22.04	<u>92.93</u>

Table 5: **VLEU scores across different ethnicities.** We sampled 1000 prompts from the visual text domain of each ethnicity using ChatGPT 3.5 and computed VLEU for several common T2I models on these prompts. Bold highlights the best score among all models, and underline underscores the best score across three races.

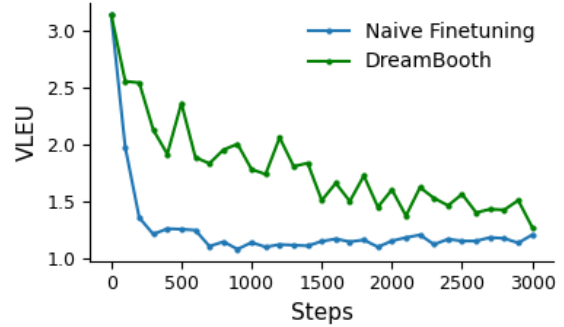


Figure 5: **VLEU of different finetuning methods.** We finetuned SD 1.5 on 5 specific teddy bears. For DreamBooth, we used 25 teddy bear images generated by the initial model as class images. During the evaluation of VLEU, we used 25 prompts about teddy bears sampled by ChatGPT 3.5.

Finetuning Methods Comparison. We compared the performance of two finetuning methods, naive finetuning and Dreambooth, on the SD 1.5 model using the DreamBooth dataset. Specifically, we select a subset of 5 teddy bear images as an example. We sampled 25 prompts about teddy bears and calculated the VLEU scores during finetuning. As expected and shown in Figure 5, Dreambooth showed a slower decline in VLEU compared to naive finetuning, aligning with its goal of preserving model generalizability during specialized training. The discernible gap in VLEU curves validates its sensitivity in capturing different rates of generalization loss. This demonstrates VLEU’s efficacy in evaluating finetuning methods’ ability to balance specificity and generalizability, a valuable asset for model development.

6 Conclusion

We introduced VLEU, a novel automatic metric to evaluate T2I models’ generalizability. VLEU quantifies alignment between sampled visual text prompts and generated images using LLMs and CLIP. Through experiments and case studies, we demonstrated VLEU’s efficacy in capturing declining generalizability during finetuning, discerning differences across models, and comparing finetuning techniques. VLEU provides an automated, standardized metric accounting for a broad space of textual prompts. Our results validate VLEU as an effective metric for quantifying T2I models’ generalizability.

7 Limitations

The efficacy of VLEU is constrained by the expressive capabilities of LLMs as the visual text sampler. Even state-of-the-art models struggle to achieve full coverage of the expansive visual text domain when sampling prompts, limiting prompt diversity for evaluation. As language models advance further in natural language understanding and generation, the accuracy and robustness of the VLEU metric will also improve.

Determining sufficient prompt quantities for robust evaluation also presents a challenge. Our experiments indicate this depends on the application scenario. For tracking declining generalizability during finetuning, relatively small samples (around 25 prompts) suffice. However, comparing generalizability across models necessitates larger samples (e.g. 1000 prompts) to ensure evaluation rigor. Further research could systematically investigate optimal prompt quantities for varying contexts to enhance VLEU stability.

While limitations exist, VLEU remains a promising metric providing a standardized framework for evaluating and improving T2I models' generalizability. Future work can explore sampling strategies and evaluation configurations to enhance VLEU robustness and utility.

8 Ethical Considerations

Our research adheres to stringent ethical standards. We utilized publicly available datasets that have been ethically vetted to avoid offensive or biased content. Participants in our human evaluation were fairly compensated, ensuring ethical treatment. Consequently, our study presents no ethical concerns, as the data is ethically sourced, the analysis unbiased, and all procedures comply with established ethical guidelines.

References

- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. [Emerging properties in self-supervised vision transformers](#). *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640.
- Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdal, and Adriana Romero-Soriano. 2021. [Instance-conditioned gan](#). In *Neural Information Processing Systems*.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. [Reproducible scaling laws for contrastive language-image learning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess Life*, 22(8):242–247.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. [An image is worth one word: Personalizing text-to-image generation using textual inversion](#). In *The Eleventh International Conference on Learning Representations*.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. [Generative adversarial networks](#). *Commun. ACM*, 63(11):139–144.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. [CLIPScore: A reference-free evaluation metric for image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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](#). In *Neural Information Processing Systems*.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [Denoising diffusion probabilistic models](#). *CoRR*, abs/2006.11239.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Tero Karras, Samuli Laine, and Timo Aila. 2019. [A style-based generator architecture for generative adversarial networks](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE.
- Alex Krizhevsky. 2012. Learning multiple layers of features from tiny images. *University of Toronto*.
- S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- Mehdi Mirza and Simon Osindero. 2014. [Conditional generative adversarial nets](#). *ArXiv*, abs/1411.1784.
- OpenAI. 2022. [Introducing chatgpt](#).
- OpenAI. 2023a. [Dall-e 3](#).
- OpenAI. 2023b. [Gpt-4](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. [SDXL: Improving latent diffusion models for high-resolution image synthesis](#). In *The Twelfth International Conference on Learning Representations*.
- 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. 2021. [Learning transferable visual models from natural language supervision](#). In *International Conference on Machine Learning*.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. [Hierarchical text-conditional image generation with clip latents](#). *ArXiv*, abs/2204.06125.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. [Zero-shot text-to-image generation](#). *CoRR*, abs/2102.12092.
- R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA. IEEE Computer Society.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. [Imagenet large scale visual recognition challenge](#). *International Journal of Computer Vision*, 115:211 – 252.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. [Photorealistic text-to-image diffusion models with deep language understanding](#). *ArXiv*, abs/2205.11487.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, page 2234–2242, Red Hook, NY, USA. Curran Associates Inc.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. [Laion-5b: An open large-scale dataset for training next generation image-text models](#). *ArXiv*, abs/2210.08402.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. [Denoising diffusion implicit models](#). *CoRR*, abs/2010.02502.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *ArXiv*, abs/2307.09288.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *Advances in neural information processing systems*, 30.

Zhonghao Wang, Wei Wei, Yang Zhao, Zhisheng Xiao, Mark A. Hasegawa-Johnson, Humphrey Shi, and Tingbo Hou. 2023. [Hifi tuner: High-fidelity subject-driven fine-tuning for diffusion models](#). *ArXiv*, abs/2312.00079.

Chenfei Wu, Sheng-Kai Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. [Visual chatgpt: Talking, drawing and editing with visual foundation models](#). *ArXiv*, abs/2303.04671.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. [Imagereward: Learning and evaluating human preferences for text-to-image generation](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Implementation Details

In this section, we provide a detailed description of the implementation process for our proposed VLEU metric. The implementation consists of two main components: sampling text prompts and calculating the VLEU score. We provide the pseudocode for each component to facilitate understanding and reproducibility.

A.1 Sampling Text Prompts

The first step in our process is to sample text prompts from the visual text domain. We use LLMs to generate these prompts. The prompts can either be random or contain a specific keyword. The pseudocode described in Algorithm 1 outlines the process of generating text prompts.

A.2 Calculating the VLEU Score

The second step involves calculating the VLEU score using the CLIP model to evaluate the semantic alignment between generated images and their corresponding text prompts. The pseudocode described in Algorithm 2 outlines the process of calculating the VLEU score.

B Human Evaluation

To demonstrate the effectiveness of our VLEU metric evaluation, we conducted a comprehensive human evaluation study involving 10 human evaluators. These evaluators were tasked with creating a variety of T2I prompts, either freely or focused on specific subjects, similar to the prompt generation process of LLMs. The prompts generated by the evaluators were then used to produce images using the four T2I models under investigation.

To facilitate the evaluation process, we developed an interactive web interface using Gradio, which allowed evaluators to compare pairs of images generated by randomly selected pairs of models. Figure 6 shows a screenshot of the Gradio interface used in the study. Evaluators were asked to determine which image better adhered to the given prompts, without knowing which model produced which image.

The pairwise comparisons made by the evaluators were used to compute an Elo rating for each model. The Elo rating system, originally developed for ranking chess players, is a method for calculating the relative skill levels of players in win-loss games. Each model’s initial rating was set to 1000. The Elo rating for a model is updated based on

Model Comparison

Constrained Subject: **Person**

Imagine a random **person** and describe he/she in one sentence. Then, compare the two images generated and select the one that better matches your description.

Figure 6: **Gradio interface used for human evaluation.** Evaluators first input a prompt related to a given subject based on the provided instructions, then click “Generate”. They then choose the better image from the two generated by different T2I models and submit their selection.

the outcome of each pairwise comparison, with the winning model gaining points and the losing model losing points. The amount of points exchanged depends on the difference in the ratings of the two models, with larger differences resulting in smaller point exchanges.

The Elo rating R for a model is updated using the following formula:

$$R_{\text{new}} = R_{\text{old}} + K \times (S - E)$$

where: - R_{new} is the new Elo rating. - R_{old} is the old Elo rating. - K is a constant that determines the sensitivity of the rating system (commonly set to 32). - S is the actual score of the match (1 for a win, 0.5 for a draw, and 0 for a loss). - E is the expected score, calculated using the formula:

$$E = \frac{1}{1 + 10^{(R_{\text{opponent}} - R_{\text{old}})/400}}$$

where R_{opponent} is the Elo rating of the opposing model.

Algorithm 1 Sampling Text Prompts

Require: number of prompts num , keyword key_word (optional), include keyword $include_key_word$ (optional), step size $step$

Ensure: A list of text prompts

```
1: Initialize LLM
2: Initialize an empty list  $prompts$ 
3: for  $i \leftarrow 0$  to  $num$  by  $step$  do
4:   if  $key\_word$  is not None then
5:     if  $include\_key\_word$  then
6:        $system\_input \leftarrow$  'Please imagine a picture of  $\{ \$key\_word \}$  and describe it
       in one sentence, making sure to include the word " $\{ \$key\_word \}$ .'"
7:     else
8:        $system\_input \leftarrow$  'Please imagine a picture of random  $\{ \$key\_word \}$  and
       describe it in one sentence.'
9:     end if
10:  else
11:     $system\_input \leftarrow$  'Please imagine a random picture and describe it in one
    sentence.'
12:  end if
13:   $human\_input \leftarrow$  [SystemMessage(content= $system\_input$ )]
14:   $ai\_output \leftarrow llm(human\_input)$ 
15:   $n \leftarrow 0$ 
16:   $limit \leftarrow \min(step, num - i)$ 
17:  while  $n < limit$  do
18:    Append AIMessage(content= $ai\_output.content$ ) to  $human\_input$ 
19:    Append HumanMessage(content='Again') to  $human\_input$ 
20:     $ai\_output \leftarrow llm(human\_input)$ 
21:    while  $key\_word$  is not None and  $include\_key\_word$  and  $key\_word$  not in
     $ai\_output.content$  do
22:       $ai\_output \leftarrow llm(human\_input)$ 
23:    end while
24:    Append  $ai\_output.content$  to  $prompts$ 
25:    Print  $ai\_output.content$ 
26:     $n \leftarrow n + 1$ 
27:  end while
28: end for
29: Return  $prompts$ 
```

Algorithm 2 Calculating the VLEU Score

Require: T2I model $t2i_model$, CLIP model $model$, T2I prompts $prompts$, number of prompts num , temperature T

Ensure: VLEU score for $t2i_model$

```
1: Load CLIP model and processor from  $model$ 
2: Load prompts from  $prompts\_path$ 
3: for  $m \in model\_types$  do
4:   Initialize  $text\_embs \leftarrow []$ ,  $img\_embs \leftarrow []$ 
5:   for  $p \in prompts$  do
6:     Tokenize  $p$ , move to device
7:     Get and normalize text features from CLIP
8:     Append to  $text\_embs$ 
9:   end for
10:  for  $p \in prompts$  do
11:    Generate image  $img$  from prompt  $p$  using  $t2i\_model$ 
12:    Move  $img$  to device, preprocess if necessary
13:    Get and normalize image features from CLIP
14:    Append to  $img\_embs$ 
15:  end for
16:  Initialize  $prob\_matrix \leftarrow []$ 
17:  for  $img\_emb \in img\_embs$  do
18:    Initialize  $cosine\_sim \leftarrow []$ 
19:    for  $text\_emb \in text\_embs$  do
20:      Calculate cosine similarity between  $img\_emb$  and  $text\_emb$ 
21:      Append to  $cosine\_sim$ 
22:    end for
23:    Calculate probability distribution using softmax with  $T$ 
24:    Append to  $prob\_matrix$ 
25:  end for
26:  Stack  $prob\_matrix$  into tensor
27:  Calculate marginal distribution for text embeddings
28:  Initialize  $image\_kl\_divergences \leftarrow []$ 
29:  for  $prob \in prob\_matrix$  do
30:    Calculate KL divergence for current image
31:    Append to  $image\_kl\_divergences$ 
32:  end for
33:  Calculate VLEU score as  $\exp(\text{avg KL divergence})$ 
34: end for
35: Return VLEU score
```

N prompts	SD 1.5	SD 2.0	SD 2.1	SDXL	N prompts	SD 1.5	SD 2.0	SD 2.1	SDXL
100	21.54	22.23	22.43	23.33	100	20.33	20.41	20.33	18.97
100	13.89	13.88	15.08	14.63	100	21.66	23.51	19.71	19.15
200	21.98	22.26	23.46	23.54	200	32.18	33.37	30.79	28.71
200	26.71	28.58	29.39	30.87	200	33.76	33.20	34.31	30.76
300	25.61	26.60	28.25	29.13	300	38.96	39.33	39.40	34.14
300	34.72	36.46	38.65	38.29	300	36.05	34.46	36.59	32.87
400	30.71	32.08	33.49	34.55	400	43.67	43.97	43.40	38.85
400	35.71	36.44	38.29	37.78	400	37.13	36.38	37.50	34.27
500	35.16	36.72	38.62	39.37	500	46.19	46.82	46.03	41.91
500	31.76	33.73	34.81	35.41	500	38.04	38.61	39.11	36.32
1000	37.18	39.15	41.78	42.05	1000	48.79	50.55	50.04	45.68
1000	29.17	30.12	32.82	32.87	1000	28.12	29.31	29.92	27.47

(a) Unconstrained subject

(b) Person-focused

Table 6: **VLEU scores for different sample sizes.** Each row represents an independent sampling. We primarily sampled two types of prompts for experimentation: unconstrained subject and person-focused prompts.

C Hyperparameter Analysis

In order to investigate the effectiveness of the VLEU metric concerning the number of sampled T2I prompts, we conducted multiple experiments across four T2I models. Each experiment involved resampling a certain number of T2I prompts using GPT 3.5 and computing the VLEU for each T2I model. The results, as shown in Table 6, indicate that as the number of sampled prompts increases, the VLEU score tends to rise. Moreover, even with sampling 1000 prompts, there is still a certain level of fluctuation in the VLEU obtained from two separate samplings. However, for a given set of sampled prompts, the relative ranking of VLEU scores among the models remains consistent. This suggests that while the absolute values of VLEU may vary between samplings, its ability to facilitate comparisons remains stable. Therefore, the choice of the number of prompts sampled could be made based on computational constraints or desired evaluation granularity, without significantly affecting the relative assessment of T2I model performance.