TYPESCORE: A Text Fidelity Metric for Text-to-Image Generative Models

Georgia Gabriela Sampaio, Ruixiang Zhang, Shuangfei Zhai, Jiatao Gu, Josh Susskind, Navdeep Jaitly, Yizhe Zhang Apple

 $\{gsamp,\ ruixiangz,\ szhai,\ jgu32,\ jsusskind,\ ndjaitly,\ yizzhang \} @apple.com$

Abstract

Evaluating text-to-image generative models remains a challenge, despite the remarkable progress being made in their overall performances. While existing metrics like CLIPScore work for coarse evaluations, they lack the sensitivity to distinguish finer differences as model performance rapidly improves. In this work, we focus on the text rendering aspect of these models, which provides a lens for evaluating a generative model's fine-grained instruction-following capabilities. To this end, we introduce a new evaluation framework called TYPESCORE to sensitively assess a model's ability to generate images with high-fidelity embedded text by following precise instructions. We argue that this text generation capability serves as a proxy for general instruction-following ability in image synthesis. TYPESCORE uses an additional image description model and leverages an ensemble dissimilarity measure between the original and extracted text to evaluate the fidelity of the rendered text. Our proposed metric demonstrates greater resolution than CLIPScore to differentiate popular image generation models across a range of instructions with diverse text styles. Our study also evaluates how well these vision-language models (VLMs) adhere to stylistic instructions, disentangling style evaluation from embedded-text fidelity. Through human evaluation studies, we quantitatively metaevaluate the effectiveness of the metric. Comprehensive analysis is conducted to explore factors such as text length, captioning models, and current progress towards human parity on this task. The framework provides insights into remaining gaps in instruction-following for image generation with embedded text.¹

1 Introduction

Image generation models have seen significant advancements in recent years, producing highquality and diverse synthetic images. Notable examples include DALL-E 3 [Ramesh et al., 2021], ideogram [Ideogram AI, 2023], Stable Diffusion [Rombach et al., 2022], MidJourney [MidJourney, 2022], Imagen [Saharia et al., 2022], Dream [WOMBO, 2021] and Adobe Firefly [Adobe, 2023]. However, even these high-quality image generation models often struggle to generate images with specific text embedded within them. This lack of *embedded-text fidelity* can take the form of typos, repeated or missing characters and words, extraneous characters, and unreadable glyphs.

Unfortunately, current metrics such as CLIPScore [Hessel et al., 2021] are unsuitable for measuring *embedded-text fidelity*. These metrics work well when there is a large performance gap between models [Bianchi et al., 2024], but are not sensitive to more nuanced improvements as the quality of generation improves [Chen et al., 2023a] because they rely on image embeddings that can lose the fine-grained details required to detect nuanced differences. As the quality of image generation has

¹Code and data will be made publicly available soon to facilitate future work on this challenging problem.

improved, there is a growing need for new metrics specifically designed to evaluate these models' instruction-following capability in a microscopic manner.



Figure 1: When assessing target image generation models p_{θ} , we provide the model with a set of instructions. These instructions prompt the model to create a set of images *i* based on specified quoted text within a particular style, alongside some contextual information. We then use a vision-language model q_{ϕ} (*e.g.* GPT-40) to extract the text from the generated images, and compute the similarity score between the generated text \hat{t} and the original quote *t*. TYPESCORE is calculated by averaging the scores obtained from multiple image generations. Common text-image alignment metrics such as CLIPScore produce indistinguishable results for both image generation models under this prompt.

We aim to bridge this gap by introducing a new evaluation metric that probes the performance differences among competitive image generation models. We propose TYPESCORE, an evaluation metric designed to assess the fidelity of embedded text in generated images (Figure 1). TYPESCORE offers a precise and nuanced assessment of embedded-text fidelity, incorporating key factors such as legibility and accuracy. Style can be a confounding factor in assessing embedded-text fidelity. We present effective methods to ground the generation with rich contextual and style instructions to create a controlled environment that minimizes the confounding aspects like text font, typeface, and aesthetic integration.

The evaluation framework for TYPESCORE includes a probing instruction dataset (TYPEINST) of 118 text-embedded image generation instructions with diverse requirements of styles, text formatting, and length. To meta-evaluate different variants of TYPESCORE and compare TYPESCORE with CLIPScore, we crowd-sourced annotations of human preferences on *text fidelity, style fidelity* and *overall preference*, over 590 pairs of generated images, and score each metric according to their alignment with the human preferences. The resulting TYPESCORE is an ensemble score over multiple dissimilarity metrics, and it aligns significantly better with human preference than CLIPScore.

We show that TYPESCORE has the sensitivity required to differentiate between the embedded-text fidelity of several state-of-the-art image generation models while CLIPScore is unable to detect these subtle differences. Interestingly, we found that models with higher TYPESCORE were also ranked higher in our annotated preferences for style-following and general instruction-following, indicating this metric can extrapolate to serve as a proxy for the model's general "instruction-following" ability. We further discuss the impact of text extraction models, instruction length, and re-captioning on the sensitivity of TYPESCORE.

While our work focuses on text fidelity evaluation in text-to-image generative models, we argue that generating faithful embedded text is actually a cutting-edge challenge that probes the most sophisticated capabilities of image generation models, thus it can potentially serve as a broader indicator of a model's fine-grained control and instruction-following abilities, as evidenced by the correlation between text fidelity scores and overall model performance in our evaluations.

Specifically, our key contributions are:

- 1. We introduce a new metric for evaluating image generation quality for rendering embedded text, highlighting the blind spots of CLIPScore. Our metric demonstrates higher sensitivity compared to existing metrics like CLIPScore for high-quality image generation models.
- 2. We provide an evaluation framework, including an instruction dataset TYPEINST with detailed and diverse image style descriptions and quoted text.
- 3. We perform large-scale human evaluation to quantitatively meta-evaluate the proposed metrics.
- 4. We provide insights into the current capabilities of image generation models and their progress towards achieving human-like proficiency in generating images with embedded text.
- 5. We conduct thorough ablation studies and analyses to understand the effects of text length, the use of different visual-language-models (VLMs) for captioning, and the impact of re-captioning.

2 Related Work

Image Generation with Embeded Text Advances in image generation models have significantly improved the quality of synthetic images. Models such as DALL-E 3 [Ramesh et al., 2021], Stable Diffusion 3 [Rombach et al., 2022], ideogram [Liu et al., 2023], and MidJourney [MidJourney, 2022] have shown remarkable progress in producing diverse and high-quality visuals. Despite their success and rapid improvement, these models still encounter challenges in generating text with high fidelity, often producing text with typos, repeated or missing characters, and extraneous glyphs. Many methods have been proposed to improve the fidelity of embedded text in the generation. TextDiffuser [Chen et al., 2023b] addresses these issues by using a two-stage process: first, a Transformer model generates the layout of keywords from text prompts, and then diffusion models generate images conditioned on these layouts. TextDiffuser-2 [Chen et al., 2023c] further enhances text rendering by integrating large language models for layout planning and text encoding, enabling more flexible and diverse text generation. AnyText [Tuo et al., 2023] takes a different approach by focusing on multilingual visual text generation and editing, leveraging a diffusion pipeline to first mask the image and then employ an Optical Character Recognition (OCR) model to encode stroke data as embeddings to generate texts that can integrate with the background. However, these models typically involve multiple components, and generating text with both high fidelity and aesthetic and natural style remains challenging, as high-fidelity text generation frequently sacrifices rendering quality and artistic value.

Text Fidelity Evaluation Metrics and Datasets Traditional text fidelity metrics such as CIDEr [Vedantam et al., 2015], SPICE [Anderson et al., 2016], and BLEU [Papineni et al., 2002a] have been widely used for evaluating image captions. CIDEr focuses on consensus in large datasets, SPICE uses scene graph structures for more detailed semantic evaluation, and BLEU measures n-gram precision against reference texts. While these metrics have been foundational, they sometimes fall short in capturing the holistic meaning. Our approach is based on these standard word-overlapping-based methods. Wang et al. [2021] introduced the FAIEr metric, designed to assess both the fidelity and adequacy of image captions. FAIEr employs a visual scene graph to bridge the image and text modalities, leveraging it as a criterion for fidelity evaluation and guiding adequacy assessment through reference captions.

When testing the instruction-following ability of image generation models, existing text-image alignment metrics like CLIPScore [Hessel et al., 2021] evaluate images by computing the cosine similarity between image and text embeddings of the instruction. Xu et al. [2023] introduced "ImageReward," a model designed to learn and evaluate human preferences for text-to-image generation, aiming to improve the alignment of generated images with human aesthetic and contextual preferences through systematic annotation and reward feedback mechanisms. Lee et al. [2024] proposes a comprehensive evaluation framework for assessing text-to-image models, addressing the limitations of existing metrics that focus narrowly on specific aspects. The framework integrates multiple dimensions, including fidelity, diversity, relevance, and creativity, to provide a more balanced assessment. The authors introduce a new benchmark combining automatic metrics and human evaluations to improve understanding and performance of these models. Somepalli et al. [2024] focuses on evaluating and understanding the stylistic attributes of images generated by diffusion models, and the proposed model demonstrates superior performance in style retrieval tasks compared to previous methods. Our approach differs from them as we focus on testing the fidelity of the embeded text in the generated image. TextDiffuser [Chen et al., 2023b] also introduces the MARIO-10M dataset and MARIO-Eval benchmark to enhance and evaluate text rendering quality. They evaluate the generation by using OCR to extract the text. Our work extensively evaluates different options for text description models and shows that OCR can yield suboptimal extraction results when the generated image is stylish.

3 Text Fidelity Assessment

Given the image generation instructions $t \sim T$, let's assume an image generation model p_{θ} produces a corresponding image $i \sim p_{\theta}(\cdot|t)$. A model exhibiting good *instruction-following* ability would create an image *i* that: 1) conforms with all the information provided in the instruction *t*, and 2) refrains from generating extraneous elements beyond the given instruction. Therefore, assessing the instruction-following capability of an image generation model can be perceived as evaluating the mutual information $\mathbf{MI}(i, t)$ of the joint distribution, which is defined as

$$\mathbf{MI}_{\theta}(i,t) = \mathbb{E}_{t \sim \mathcal{T}, i \sim p_{\theta}(\cdot|t)} \log \frac{p(i,t)}{p(i)p(t)}$$
(1)

However, directly evaluating the instruction-following capability would be challenging as there are numerous ways to describe an image. To probe the general instruction-following capability of an image generation model, we study a more *controlled* problem of the *embedded-text fidelity assessment* task, which evaluates how faithful an image generation model follows the instruction to render a specific piece of text in a certain style. As the evaluation focuses on the embedded-text fidelity, this yields a clear evaluation metric.

Consider a dataset \mathcal{D} containing N image generation instructions, where each instruction includes a quoted text t. Each image generation model is tasked with generating an set of images $\{i_1, \dots, i_N\}$, based on the instructions. We investigate the following: how accurately are the rendered text in the images compared to the quoted text t from the instruction?

3.1 **TYPESCORE: A Text Fidelity Evaluation Framework for Image Generation Models**

Suppose we have a reverse model q_{ϕ} that can predict the instruction t from the image i, from (1) (see Appendix A for proof),

$$\mathbf{MI}_{\theta}(i,t) \ge \mathbb{E}_{t \sim \mathcal{T}} \mathbb{E}_{i \sim p_{\theta}(\cdot|t)} \log q_{\phi}(t|i) \triangleq \mathcal{L}_{\mathbf{MI}}(\theta;\phi).$$
⁽²⁾

This suggests that instead of directly estimating **MI**, we can potentially use an image description model q_{ϕ} to calculate a lower bound proxy $\mathcal{L}_{\mathbf{MI}}(\theta; \phi)$ of **MI**. Due to the rapid advancement of Vision-Language Models (VLMs), obtaining this q_{ϕ} becomes more convenient and q_{ϕ} can be good offthe-shelf zero-shot posterior approximators in this context. To evaluate the image generation model's capability in generating accurate text based on instructions, we introduce an evaluation framework that leverages an image description model q_{ϕ} . Practically, we can ask q_{ϕ} to either calculate the likelihood of t, or to generate an estimate \hat{t} , using a similarity measure $S(t, \hat{t})$ as the metric when q_{ϕ} does not produce a likelihood.

In the following, we discuss our evaluation framework, which consists of a dataset of diverse instructions, an image description model to extract text from the images, and an ensemble score to measure the difference between the extracted text and the original quoted text.

3.2 TYPEINST Dataset

We create the dataset of text instructions, TYPEINST, using GPT-3.5 [Brown et al., 2020] by prompting the model with basic text and style elements written by the authors. Following the Magic Prompt approach in ideogram [Ideogram AI, 2023], the GPT-3.5 model was prompted to enhance and recaption the initial raw descriptions in three iterations, resulting in rich instructions that offer comprehensive details about both the image and the text style. The text to be rendered is in between quotes. See Figure 2 for sampled generations using the instructions from TYPEINST.



Figure 2: Sampled generations of ideogram using the instructions from TYPEINST.

TYPEINST dataset comprises 118 instructions across various scenarios for evaluation. The average number of words per instruction is 33.94, while the average number of words in the quote is 11.78. TYPEINST covers a broad range of domains and subjects (see Appendix B for the composition of TYPEINST) such as celebratory milestones, futuristic adventures, urban life, cozy settings, inspirational messages, historical themes, cultural celebrations, natural landscapes, educational environments, and artistic expressions. It also includes practical text instances beyond English characters, like addresses, digits, acronyms, and logos. This variety encompasses different styles, fonts, and contexts, providing a comprehensive resource for evaluation.

3.3 Image Description Methods

With the generated images from each tested model, we employ an image description model q_{ϕ} to extract the rendered text from these images. We compared the performance of two different classes of image captioning models to extract captions: OCR [Jana et al., 2014], and Vision-language Models (VLMs). The VLMs were instructed to extract only the rendered text while preserving any typos or errors. We also ask the model to use "@" tokens to represent any glyphs or extraneous symbols that cannot be reasonably interpreted to match any English character (Figure 1). The full prompt is provided in below:

Identify the main text contained in this image, and output it between quotes, without correcting any typos or issues you may encounter. **Do not output anything else.**

To compare these models, we ask human annotators to extract the rendered text from 590 generated images (see Appendix D for details and the screenshot of the annotation interface). Using this human extraction as ground truth \hat{t}_{oracle} , we can compute the extraction accuracy using Normalized Edit distance (NED) [Yujian and Bo, 2007] for each of the q_{ϕ} . Note that as some of the q_{ϕ} tend to auto-correct the extracted text, the similarity score between original text t and the extracted text $\hat{t} \sim q_{\phi}(\cdot|i)$ can be even higher than the similarity between t and \hat{t}_{oracle} . Therefore, the automatic

score of similarity cannot be used to judge on which q_{ϕ} is more accurate. The comparison of the q_{ϕ} is provided in Table 1.

We observed that OCR underestimates the TYPE-SCORE by failing to identify the main text and introducing extraneous characters and symbols from glyphs. Conversely, VLMs tend to overestimate TYPESCORE by fixing typos and incorrect word ordering 3. However, with a careful prompt tuning, this overestimation issue can be alleviated. In practice, we used GPT-40 [OpenAI, 2024] as in our experiments it leads to the best extraction accuracy comparing to other alternatives, including OCR, GPT-4v and LLaVa-NeXT [Liu et al., 2024]. We combined OCR and GPT-40 (**OCR+GPT-40**) by feeding the OCR output into GPT-40 and prompting the model to discern which portions of the

Models	$\text{NED}(\hat{t}_{\text{oracle}}, \hat{t}_{q_{\phi}}) (\downarrow)$
OCR	0.650 ± 0.032
LLaVa-NeXT	0.618 ± 0.023
GPT-4v	0.340 ± 0.029
OCR + GPT-40	0.355 ± 0.033
GPT-40	$\textbf{0.315} \pm \textbf{0.030}$

Table 1: The Normalized Edit Distance between human oracle extraction and each q_{ϕ} 's extraction. GPT-40 yields the highest alignment with human oracle extraction.

OCR output were extraneous to the main text, thereby preserving the original typos from OCR while utilizing GPT-4o's strength in accurately identifying the main text. Combining OCR and GPT-4o fails to outperform GPT-4o. The prompt used to refine the OCR output with GPT-4o is provided in Appendix C.1.



Generated image i

Extracted text \hat{t}

Figure 3: When extracting text, OCR tends to introduce errors, while VLMs tend to autocorrect existing errors in the rendered text.

3.4 Scoring Mechanism

Image generation models may struggle to accurately render text due to various factors, such as typos (missing, repeated, or unnecessary characters), errors in word order or repetition, and even generating unintelligible text or no text at all. Given the extracted text \hat{t} , we explored a set of dissimilarity metrics that cover different error factors.

- **Normalized Edit Distance** [Yujian and Bo, 2007] measures the ratio of edit distance to the average of the length of both strings, providing a normalized measure of dissimilarity between two strings. These metrics were chosen for their effectiveness in quantifying deviations at the character level, enabling the identification of typographical anomalies such as misspellings, character omissions, and insertions.
- **BLEU** [Papineni et al., 2002b] evaluates the precision of machine-generated translations by comparing them to a reference translation based on exact word matches. BLEU-1 was selected to evaluate fidelity at the word level, enabling the detection of discrepancies in word choice, repetition, and omission.
- **Character-BLEU** evaluates the precision of machine-generated translations by comparing them to a reference translation based on exact *character matches*, rather than entire words.
- Normalized Longest Common Subsequence (NLCS) [Ullman et al., 1976] measures the length of the longest common subsequence between two strings, normalized by dividing

by the length of the longer string, providing a similarity measure between the two strings. These metrics were selected to assess text completeness by quantifying the extent to which the generated text aligns with the original text, thereby providing insights into both word order fidelity and text integrity.

- Smith Waterman [Smith and Waterman, 1981] is a local sequence alignment algorithm used to identify the optimal local alignment between two sequences by scoring matches, mismatches, and gaps.
- Ensemble Score comprises a subset of the aforementioned distance metrics.

Alternatively, we could compute the likelihood $q_{\phi}(t|i)$ for the original quote t. Directly calculating such likelihood by extracting all logits from the GPT-4 API is challenging because the API does not natively support log probability evaluation of input tokens. It only supports log probability evaluations for up to the top 20 generated tokens, making it difficult to compute likelihoods in a reasonable and reliable manner.

3.5 Evaluated Baseline Models

We evaluated four image generation models that showcase the SOTA text rendering capabilities of current VLMs: DALL-E 3 [Ramesh et al., 2021], Stable Diffusion XL [Podell et al., 2023], Stable Diffusion 3 [Stability.ai, 2024], and ideogram [Ideogram AI, 2023]. These models were used only for inference in their default configurations to generate images according to the given instructions. We use 8x Nvidia A100 GPUs for all the experiments.

4 Meta-evaluation of the TYPESCORE Variants

Human annotation To meta-evaluate different variants of TYPESCORE and compare our method with CLIPScore, we further performed pairwise human evaluation task on 472 pairs of images generations from DALL-E 3, ideogram, Stable Diffusion and Stable Diffusion 3 on our internal crowd-source annotation platform. Each pair of generated images was evaluated by 3 to 5 judges, presented in random order with in-context examples and detailed instructions of the requirements and evaluation aspects. We aimed for at least 60% agreement: if 2 out of 3 judges concurred on a top answer, the evaluation concluded. Otherwise, up to 2 additional judges were consulted to achieve the desired agreement. The judges were instructed to assess on three tasks using a 3-point Likert-like scale:

- 1. **Text fidelity:** *In which image does content of the rendered text better align with the original quote?*
- 2. **Style fidelity:** *In which image does the style better align with the style description in the instruction?*
- 3. **Overall preference:** *Considering the content of the rendered text, alignment with the instruction and aesthetic value, which image better aligns with the given instruction?*

Further details, including the human evaluation template used, hourly rate of the evaluation task, and inter-rater agreement analysis, are provided in Appendix D.

Results Based on the **Text fidelity** annotation, we compare different variants of TYPESCORE via the meta-metrics of *alignment accuracy*, which is computed as

Alignment Accuracy =
$$\frac{|(\text{TYPESCORE}(\theta) > \text{TYPESCORE}(\theta')) \cap \text{Human prefer } \theta \text{ over } \theta'|}{|\text{All annotated pairs}|}$$

where $|\cdot|$ denotes the cardinality. The alignment accuracy indicates the percentage of pairs where automatic metrics and human preferences agree. Tied pairs are excluded from the calculation. This metric is linked to Percent Rank Violation (PRV) [Li et al., 2024], which evaluates the ranking violation of the metrics against the oracle preference. In this way, we calculate the ranking violation of the metrics for each pair of models using human-annotated data, then compute the aggregated means. The results are presented in Table 2. We found that TYPESCORE (character-BLEU), TYPESCORE (Smith Waterman) and TYPESCORE (NED) generally exhibit better alignment with human judgement of text

fidelity, style fidelity and overall preference compared to other distance metrics. After normalizing each distance metric to [0, 1], combining NED, Smith-Waterman and NLCS through mean pooling effectively leverages their strengths, resulting in robust, generalizable and high alignment accuracy. We refer to the resulting ensemble methods as TYPESCORE in the subsequent discussion. TYPESCORE consistently outperforms CLIPScore in this meta-evaluation. Further examples are provided in the Appendix F, illustrating how TYPESCORE can discern subtle differences in rendered text. We also noted that in cases where there is a significant quality gap between pairs, both CLIPScore and TYPE-SCORE can accurately detect the difference. Yet, when the quality gap is narrower (*e.g.*, DALL-E vs Ideogram), CLIPScore consistently struggles to rank them correctly, whereas TYPESCORE remains sensitive.

Alignment Accuracy	Text Fidelity(\uparrow)	Style Fidelity(†)	Overall Prefer. (†)
CLIPScore [Hessel et al., 2021]	$66.3\% \pm 0.6\%$	$58.3\% \pm 0.9\%$	$65.8\% \pm 0.7\%$
TYPESCORE (NED)	$69.0\% \pm 0.5\%$	$61.7\% \pm 0.5\%$	$68.6\% \pm 0.3\%$
TYPESCORE (BLEU)	$69.3\% \pm 0.4\%$	$60.6\% \pm 0.9\%$	$69.2\% \pm 0.5\%$
TYPESCORE (character-BLEU)	$\mathbf{71.5\%}\pm0.8\%$	$62.0\% \pm 0.5\%$	$71.1\% \pm 0.7\%$
TYPESCORE (NLCS)	$67.8\% \pm 0.6\%$	$59.3\% \pm 0.9\%$	$67.7\% \pm 0.8\%$
TYPESCORE (Smith Waterman)	$69.3\% \pm 0.5\%$	$63.9\% \pm 0.7\%$	$69.2\% \pm 0.4\%$
TYPESCORE (Ensemble Score)	$71.1\% \pm 0.5\%$	$62.2\% \pm 0.7\%$	$\textbf{71.3\%} \pm 0.3\%$

Table 2: Alignment Accuracy of CLIPScore and TYPESCORE variants based on human preference, w.r.t text fidelity, style fidelity and overall preference. GPT-40 is used for the text extraction. TYPESCORE aligns better with human preference of text fidelity, style fidelity and overall preference. Averaging the three columns, TYPESCORE (Ensemble Score) yields the robust and highest alignment accuracy.

Extrapolation property of TYPESCORE Interestingly, as shown in Table 2, our approach also demonstrates good alignment with **Style fidelity** and **Overall preference**, while CLIPScore falls short. This suggests that text fidelity can be closely associated with the evaluation of general instruction-following ability of an image generation model. Therefore, we can use TYPESCORE to *probe* the image generation model's capability to follow instructions, particularly if they are not specifically tailored to optimize rendered text fidelity.

TYPESCORE sensitivity to Text Length We found that TYPESCORE is robust to variations in input text length, yielding consistent performance regardless of the length of the instruction or quoted text. We validated this robustness by computing the Pearson correlation [Pearson, 1895, Stigler, 1989] between TYPESCORE and varying input text lengths, and found no significant correlation across several image generation models. This indicates that the TYPESCORE is stable and reliable across texts of varying lengths. See Appendix E.1 for more details.

TYPESCORE sensitivity to Recaptioning We found that recaptioning the input image description by adding more stylistic details and contextual information helps control the text fidelity evaluation of the rendered text. This improvement is reflected in higher TYPESCORE and slightly lower variance, demonstrating that incorporating richer stylistic nuances contributes to a more controlled setting for text fidelity evaluation. We measured this by prompting ideogram to generate images using an augmented input caption and comparing the scores of the resulting images. See Appendix E.2 for more details.

TYPESCORE sensitivity to Text Extraction Method We conducted a comprehensive analysis to evaluate TYPESCORE's robustness across different text extraction methods. While our primary quantitative analyses utilize GPT-40 due to its highest alignment with human extractions, we found that TYPESCORE maintains consistent performance across different text extraction approaches. To demonstrate this, we evaluated the alignment accuracy using LLaVA-NEXT, a more efficient alternative to GPT-40. Despite LLaVA-NEXT showing lower alignment with human oracle extractions (0.618 mean normalized edit distance) compared to GPT-40, TYPESCORE consistently outperforms CLIPScore in aligning with human preferences across all evaluation aspects (Table 3).

Alignment Accuracy (LLaVA):	Text Fidelity(\uparrow)	Style Fidelity(†)	Overall Prefer. (†)
CLIPScore	$38.4\% \pm 0.8\%$	$36.1\% \pm 0.5\%$	$38.0\% \pm 0.5\%$
TYPESCORE (NED)	$39.9\% \pm 0.4\%$	$37.1\% \pm 0.9\%$	$39.9\% \pm 0.7\%$
TYPESCORE (BLEU)	$38.9\% \pm 0.5\%$	$\mathbf{37.7\%}\pm0.6\%$	$38.9\% \pm 0.5\%$
TYPESCORE (character-BLEU)	$\textbf{40.2\%}\pm0.6\%$	$37.5\% \pm 0.9\%$	$\textbf{40.2\%}\pm0.5\%$
TYPESCORE (NLCS)	$38.9\% \pm 0.3\%$	$36.5\% \pm 0.8\%$	$38.7\% \pm 0.6\%$
TYPESCORE (Smith Waterman)	$39.9\% \pm 0.4\%$	$\textbf{37.7\%} \pm 0.7\%$	$40.0\% \pm 0.7\%$
TYPESCORE (Ensemble Score)	$39.0\% \pm 0.4\%$	$37.0\% \pm 0.7\%$	$39.3\% \pm 0.6\%$

Table 3: Alignment accuracy using LLaVA-NEXT as the text extraction model. Despite using a less accurate text extraction model, TYPESCORE maintains better alignment with human preferences compared to CLIPScore across all evaluation aspects.

This robustness can be attributed to the fact that a weaker text extraction model tends to introduce similar levels of errors across all generated images being compared, thus not significantly affecting their relative rankings in the evaluation. This flexibility allows users to employ any state-of-the-art text extraction model based on their specific requirements and constraints, while maintaining reliable evaluation results.

5 Evaluation of Image Generation Models using TYPESCORE

Tested Model	TYPESCORE([†])	Style Fidelity(†)	Overall Preference ([†])
Stable Diffusion XL	0.238 ± 0.013	0.25	0.02
DALL-E 3	0.739 ± 0.018	0.87	0.54
Stable Diffusion 3	0.800 ± 0.016	0.73	0.17
ideogram	$\textbf{0.882} \pm \textbf{0.009}$	0.87	0.70

Table 4: Evaluation of several image generation models using TYPESCORE. Ideogram outperforms the others models under TYPESCORE. It also garners the top human rating for **style fidelity** and **overall preference**.

We assess the image generation models mentioned in section 3.5 using TYPESCORE. The results are presented in Table 4. Our evaluation indicates that ideogram attained the highest TYPESCORE, with stable diffusion 3 coming in second. Despite DALL-E 3's capability to generate high-quality images, it falls short in accurately rendering the text accurately.

6 Limitations

Despite the promising results of TYPESCORE in evaluating text fidelity in synthetic images, several limitations must be acknowledged. First, the reliance on existing VLMs for text extraction introduces dependency on their performance and limitations. In scenarios where the VLMs themselves exhibit biases or inaccuracies, these will propagate into our evaluation, potentially skewing TYPESCORE results. Moreover, the diverse nature of text styles and formats in TYPEINST may not comprehensively cover all real-world use cases. Our metric is evaluated on Latin text and should benefit from being evaluated with non-Latin text as well. Lastly, our human evaluation process, although extensive, is subject to individual annotator biases and interpretations. While we have taken measures to ensure consistency and reliability, human evaluations inherently carry a degree of subjectivity that can influence the assessment outcomes.

7 Conclusion

We introduced a comprehensive evaluation framework, TYPESCORE, designed to assess the fidelity of text embedded in synthetic images generated by various models. Our framework evaluates the degree to which the generated images accurately follow textual instructions, leveraging a combination

of automatic metrics based on human judgment. By comparing the performance of our metrics with human preferences, we demonstrated that TYPESCORE aligns more closely with human judgment compared to traditional metrics like CLIPScore, indicating its efficacy for evaluating text fidelity in image generation models. In future work, we aim to explore the possibility of extending our approach to evaluate image generation in general domains beyond text rendering. We also plan to assess whether calculating the likelihood $q_{\phi}(t|i)$ could provide a more precise evaluation of text fidelity compared to the dissimilarity metrics we used in TYPESCORE.

8 Acknowledgement

We would like to thank Ziv Wolkowicki, Barry Theobald, Samy Bengio, Richard Bai, Zijin Gu, and Tatiana Likhomanenko for their valuable feedback and contributions.

References

- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. URL https://arxiv.org/abs/2102.12092.
- Ideogram AI. Ideogram: Ai for image and text generation. http://ideogram.ai, 2023.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Highresolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022. URL https://arxiv.org/abs/2112.10752.

MidJourney. Midjourney: Ai-based art generation. https://www.midjourney.com, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/ hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

WOMBO. Dream. https://dream.ai, 2021.

Adobe. Adobe firefly. https://www.adobe.com/products/firefly.html, 2023.

- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A referencefree evaluation metric for image captioning. *ArXiv*, abs/2104.08718, 2021. URL https://api. semanticscholar.org/CorpusID:233296711.
- Lorenzo Bianchi, Fabio Carrara, Nicola Messina, and Fabrizio Falchi. Is CLIP the main roadblock for fine-grained open-world perception? *CoRR*, abs/2404.03539, 2024. doi: 10.48550/ARXIV. 2404.03539. URL https://doi.org/10.48550/arXiv.2404.03539.
- Cangxiong Chen, Vinay P. Namboodiri, and Julian Padget. Understanding the vulnerability of clip to image compression, 2023a.
- Ming Liu, Fang Wu, and Lei Shen. Ideogram: Simplifying text-to-image generation for everyone. arXiv preprint arXiv:2301.09876, 2023. URL http://ideogram.ai.
- Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. In *NeurIPS 2023*, May 2023b. URL https://www.microsoft.com/en-us/research/publication/ textdiffuser-diffusion-models-as-text-painters/.
- Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. *arXiv preprint arXiv:2311.16465*, 2023c.

- Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023. URL https://arxiv.org/abs/2311.03054.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002a.
- Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059, 2021.
- 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. arXiv preprint arXiv:2304.05977, 2023. URL https://arxiv.org/abs/2304.05977.
- Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- Ranjan Jana, Amrita Chowdhury, and Sk Islam. Optical character recognition from text image. International Journal of Computer Applications Technology and Research, 3:240–244, 04 2014. doi: 10.7753/IJCATR0304.1009.
- Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007. doi: 10.1109/TPAMI.2007.1078.
- OpenAI. Gpt-40. https://openai.com/index/hello-gpt-40/, 2024.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl. github.io/blog/2024-01-30-llava-next/.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Annual Meeting of the Association for Computational Linguistics, 2002b. URL https://api.semanticscholar.org/CorpusID:11080756.
- J. D. Ullman, A. V. Aho, and D. S. Hirschberg. Bounds on the complexity of the longest common subsequence problem. *J. ACM*, 23(1):1–12, jan 1976. ISSN 0004-5411. doi: 10.1145/321921. 321922. URL https://doi.org/10.1145/321921.321922.
- Temple F. Smith and Michael S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147 1:195–7, 1981. URL https://api.semanticscholar.org/ CorpusID:20031248.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

Stability.ai. Stable diffusion 3. https://stability.ai/news/stable-diffusion-3, 2024.

- Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv* preprint arXiv:2405.05941, 2024.
- Karl Pearson. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242, January 1895.
- Stephen M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2): 73–79, 1989. ISSN 08834237. URL http://www.jstor.org/stable/2245329.

Appendix

A **Proof of** (1)

$$\mathbf{MI}_{\theta}(i,t) = \mathbb{E}_{t \sim \mathcal{T}, i \sim p_{\theta}(\cdot|t)} \log \frac{p(i,t)}{p(i)p(t)}$$

= $H(t) + \mathbb{E}_{i}D_{KL}(p_{\theta}(t|i), q_{\phi}(t|i)) + \mathbb{E}_{t \sim \mathcal{T}, i \sim p_{\theta}(\cdot|t)} \log q_{\phi}(t|i)$
 $\geq \mathbb{E}_{t \sim \mathcal{T}}\mathbb{E}_{i \sim p_{\theta}(\cdot|t)} \log q_{\phi}(t|i) \triangleq \mathcal{L}_{\mathbf{MI}}(\theta;\phi).$

where $H(\cdot)$ represents the entropy. D_{KL} denotes the KL divergence between two distributions.

B Dataset composition

We provide the dataset composition of TYPEINST below. TYPESCORE covers a variety of scenarios that instruct the image generation model to render text.



Figure 4: Composition of TYPEINST dataset.

C Image description methods

C.1 OCR+GPT-40 prompt

When combining **OCR+GPT-40**, we used the prompt below to generate a response. In the prompt. {ocr_extracted_caption} refers to the caption extracted by OCR.

This image contains a main quote and it might contain additional text. We already extracted both the main quote and any additional text from the image, and it follows: {ocr_extracted_caption}. We want to isolate only the main quote. From this text, identify the main quote and extract it in the right order, without correcting any typos or issues you may encounter, and without adding any new words. Output the main quote between quotes and do not output anything else.

C.2 GPT-40 versus LLaVa

We assess the image generation models mentioned in section 3.5 using TYPESCORE with LLaVA-NeXT. The results are presented in Table 5. As an open-source model, LLaVa offers lower costs compared to GPT-40, making it a cost-effective alternative for TYPESCORE.

Tested Model	TYPESCORE (GPT-40)	TYPESCORE (LLaVa)
Stable Diffusion XL	0.238 ± 0.013	0.357 ± 0.088
DALL-E 3	0.739 ± 0.018	0.746 ± 0.018
Stable Diffusion 3	0.800 ± 0.016	0.816 ± 0.017
ideogram	$\textbf{0.882} \pm \textbf{0.009}$	0.855 ± 0.013

Table 5: Evaluation of several image generation models using TYPESCORE with LLaVa. Similarly to TYPESCORE with GPT-40, ideogram outperforms the other models under TYPESCORE with LLaVa.

D Human evaluation details

In total, we recruited 104 human annotators to participate in the study using our internal crowd-source annotation platform. Annotators were recruited from Canada, Great Britain, the United States, Australia, Singapore and India and paid an average of \$76.28 USD per hour.

To ensure the quality of the annotation data, we qualified each rater by assigning a set of test questions for which the answers were known. In order to be qualified to annotate the study, raters were required to correctly answer at least 90% of the test set questions. Additionally, we manually inspected the annotations to validate the human ratings. We found that the inter-rater agreement was high, with only 4.87% of the tasks requiring two additional judges to reach an agreement.

We show the UI for the annotation task in Figures 5 and 6: Figure 5 shows the UI for the tutorial that was provided to each annotator, and Figure 6 shows the task UI.

D.1 Tutorial UI

Instructions:

We're comparing different methods for generating images from a given image description. The image description contains a quote as well as a style description explaining the scene in which the text appears.

In this task, you'll see two images and your job is to pick the image that best matches the description, with respect to the content of the text and the style of the image.

Section 1 - Content of the text:

Instruction: Select the image in which the content of the text best matches the quote from the description.

There is some room for subjectivity here, since there could be several types of errors in the images, like repeated or missing characters, repeated or missing words, and even unreadable text. Please select the best image using the following guideline:

Best images: The best images are the images that have the fewer number of errors.

Worst images: The worst type of errors is unreadable text. If an image has no text at all or contains only scribbles that look like text but are unreadable, that image is automatically considered worse.

Breaking ties: For a pair of images that have the same number of errors, break ties by giving preferences to errors that occur at the character level, followed by errors that occur at the word level.

Errors at the character level: missing characters / repeated characters / wrong order of characters

Errors at the word level: missing words / repeated words / wrong order of words

Example 1:

Image description: A white coffee mug with elegant handwriting that says "Good Morning, Sunshine!" Text is horizontal, centered, using a playful font, medium size, minimalist art style. The mug is placed on a wooden table next to a cozy window with morning light streaming in, casting a warm glow on the scene.



QUESTION: In which image does the content of the text best match the quote from the description?

Anowen. intage i.

EXPLANATION: While Image 2 contains errors at the character level, it has 4 repeated characters, while Image 1 has only 1 repeated word. 4 is greater than 1, so we pick image 1.

Example 2:

Image description: A white coffee mug with elegant handwriting that says "Good Morning, Sunshine!" Text is horizontal, centered, using a playful font, medium size, minimalist art style. The mug is placed on a wooden table next to a cozy window with morning light streaming in, casting a warm glow on the scene.

Image 1

mage 2



QUESTION: In which image does the content of the text best match the quote from the description? ANSWER: Image 1.

EXPLANATION: While Image 1 contains 4 repeated characters, but Image 2 only has unreadable scribbles.

Section 2 - Style of the image:

Instruction: Select the image in which the style of the image best matches the description.

In this section, please completely ignore the content of the text, and consider only the style of the font and the background image.

The best images are the images that best match the style description, with respect to the text font, color, style, textures, lighthing and background elements.

Example 3:

A white coffee mug with elegant handwriting that says "Good Morning, Sunshine!" Text is horizontal, centered, using a playful font, medium size, minimalist art style. The mug is placed on a wooden table next to a cozy window with morning light streaming in, casting a warm glow on the scene.



QUESTION: Does the style in IMAGE 1 match the style from the description?

ANSWER: Yes

EXPLANATION: The image matches the style description. There is a white coffee mug with text written in elegant handwriting, positioned at the center of the image, in minimalist art style. The mug is placed on a wooden table next to a window with morning light streaming in.

QUESTION: Does the style in IMAGE 2 match the style from the description?

ANSWER: Yes

EXPLANATION: Similarly to Image 1, Image 2 matches the style description. There is a white coffee mug with text written in elegant handwriting, positioned at the center of the image, in minimalist art style. The mug is placed on a wooden table next to a window with morning light streaming in.

QUESTION: Considering only the image style, which image do you prefer?

ANSWER: Image 2

EXPLANATION: Both images match the style description, but the mug in image 1 has two handles, which was not specified in the image description.

Section 3 - Which image best matches the image description?

Instruction: Consider both the content of the text and the style of the image, and select the image that best matches the image description.

There is some room for subjectivity here because the images may have different levels of instruction following for the text description and the image description. In these cases, choose the image that best matches the image description overall.

If you think that both images are equally good (or equally bad) at matching the description, indicate it by selecting "Equally".

Figure 5: The tutorial UI is split into 3 sections, corresponding to the sections of the annotation task: text fidelity, style fidelity, and overall preference. Each section contains example image pairs that demonstrate potential issues annotators might encounter, along with the correct answers for each scenario.

D.2 Annotation Task UI

Image description:

A cake with writing on it that says "congratulations Kate and Luke on your upcoming arrival". The text is delicately scripted in a cursive font with a slight tilt angle, gracefully embellished with floral motifs. The size of the text covers a quarter of the cake's surface, elegantly presented in a calligraphic style.

Images:



Section 1 - Content of the text:

Input quote: congratulations kate and luke on your upcoming arrival

CONTENT OF THE TEXT: In which image does the content of the text best match the quote from the description?

- Image 1
- O Image 2
- Equally

Section 2 - Style of the image:

Style description: A cake with writing on it that says (quote). The text is delicately scripted in a cursive font with a slight tilt angle, gracefully embellished with floral motifs. The size of the text covers a quarter of the cake's surface, elegantly presented in a calligraphic style.

STYLE IMAGE 1: Does the image style in IMAGE 1 follow the style from the instruction?

YesNo

STYLE IMAGE 2: Does the image style in IMAGE 2 follow the style from the instruction?

YesNo

STYLE PREFERENCE: Considering only the image style, which image do you prefer?

Image 1
 Image 2

Equally

Image 1
 Image 2
 Equally

Section 3 - Which image best matches the image description?

Image description: A cake with writing on it that says "congratulations Kate and Luke on your upcoming arrival". The text is delicately scripted in a cursive font with a slight tilt angle, gracefully embellished with floral motifs. The size of the text covers a quarter of the cake's surface, elegantly presented in a calligraphic style.

OVERALL ASSESSMENT: Given the instruction provided, which image best matches the image description?

Figure 6: Annotation task UI. Users we provided with an image description and a pair of images, and asked to rate the images with respect to their text fidelity, style fidelity, and overall preference.

٥

٥

n

٥

0

E TYPESCORE Sensitivities

E.1 Sensitivity to the length of the instruction

Table 6 shows the Pearson correlation coefficients between different variants of TYPESCORE and overall input caption lengths across different image generation models. This suggests that TYPE-SCORE exhibits minimal correlation with caption length for each model, indicating it yields stable score across various lengths of the input text.

TYPESCORE variants:	SD XL	DALL-E 3	SD 3	ideogram
TYPESCORE (NED)	0.00	0.01	0.00	0.00
TYPESCORE (BLEU)	0.06	0.03	0.12	0.04
TYPESCORE (character-BLEU)	0.03	0.04	0.02	0.07
TYPESCORE (NLCS)	0.04	0.04	0.02	0.05
TYPESCORE (Smith Waterman)	0.10	0.02	0.03	0.03
TYPESCORE (Ensemble Score)	0.03	0.01	0.02	0.02

Table 6: Pearson correlation coefficients between TYPESCORE and different input caption lengths. SD denotes Stable Diffusion. The results show no significant correlation across various image generation models, demonstrating that TYPESCORE remains stable across various text lengths.

E.2 Sensitivity to text recaptioning

We found that augmenting the input instruction with more detailed stylistic details and contextual information helps control the text fidelity evaluation of the rendered text. The results are shown in Table 7. We assessed this by comparing the quantitative results of ideogram and ideogram Magic Prompt E.2. The ideogram Magic Prompt model is an extension of the default ideogram model, where an augmented image instruction is suggested and used to generate the image 7.

It can be observed that TYPESCORE has a higher mean value and slightly lower variance, demonstrating that incorporating richer stylistic nuances contributes to a more controlled setting for text fidelity evaluation.

TYPESCORE variants:	ideogram	ideogram Magic Prompt
TYPESCORE (NED) (\downarrow)	0.138 ± 0.013	$\textbf{0.124} \pm \textbf{0.014}$
TYPESCORE (BLEU) (\uparrow)	0.691 ± 0.018	0.772 ± 0.015
TYPESCORE (character-BLEU) (\uparrow)	0.878 ± 0.010	$\textbf{0.912} \pm \textbf{0.009}$
TYPESCORE (NLCS) (\uparrow)	0.924 ± 0.006	0.936 ± 0.006
TYPESCORE (Smith Waterman) (\uparrow)	0.899 ± 0.009	0.909 ± 0.009
TYPESCORE (Ensemble Score) (\uparrow)	0.882 ± 0.009	0.895 ± 0.009

Table 7: ideogram with recaptioning consistently outperforms ideogram with standard prompts.

F Comparative qualitative analysis via example generations

In the following sections, we present example generations from each image generation model. Each model's outputs are divided into two columns: left and right. The left column showcases the most faithful generations from the model, while the right column displays examples with the lowest text fidelity. These low-fidelity examples often feature numerous typos, repeated words and characters, illegible glyphs, or a complete absence of text. These comparisons help elucidate the differences in text fidelity among the models.



Figure 7: The ideogram Magic Prompt model augments the contextual information of the input instruction, adding more stylistic details. We found that this helps control the text fidelity evaluation of the rendered text.

F.1 Stable Diffusion XL



Table 8: Stable Diffusion XL example generations. On the **left**, see some generations with higher text fidelity. On the **right**, see some generations with lower text fidelity.

F.2 DALLE 3



Table 9: DALLE-3 example generations. On the **left**, see some generations with higher text fidelity. On the **right**, see some generations with lower text fidelity.

F.3 Stable Diffusion 3



Table 10: Stable Diffusion 3 example generations. On the **left**, see some generations with higher text fidelity. On the **right**, see some generations with lower text fidelity.

F.4 ideogram



Table 11: ideogram example generations. On the **left**, see some generations with higher text fidelity. On the **right**, see some generations with lower text fidelity.