TextTIGER: Text-based Intelligent Generation with Entity Prompt Refinement for Text-to-Image Generation

Anonymous ACL submission

Abstract

Generating images from prompts containing specific entities requires models to retain as much entity-specific knowledge as possible. However, fully memorizing such knowledge is impractical due to the vast number of entities and their continuous emergence. To address this, we propose Text-based Intelligent Generation with Entity prompt Refinement (TextTIGER), which augments knowledge on entities included in the prompts and then summarizes the augmented descriptions using Large Language Models (LLMs) to mitigate performance degradation from longer inputs. To evaluate our method, we introduce WiT-Cub (WiT with Captions and Uncomplicated Background-explanations), a dataset comprising captions, images, and an entity list. Experiments on multiple image generation models and LLMs show that TextTIGER improves image generation performance in standard metrics (IS, FID, and CLIPScore) compared to caption-only prompts. Additionally, multiple annotators' evaluation confirms that the summarized descriptions are more informative, validating LLMs' ability to generate concise yet rich descriptions. These findings demonstrate that refining prompts with augmented and summarized entity-related descriptions significantly enhances image generation capabilities. The dataset will be available upon acceptance.

1 Introduction

Text-to-Image is a task to generate images from given texts. To convert textual information into an image, image generation models such as Stable Diffusion (Rombach et al., 2022) rely on a diffusion model (Ho et al., 2020) with a text encoder, which requires precise and appropriate prompts that capture the images they intend to generate. In this process, the image generation models should retain as much entity-specific knowledge, e.g., the names of buildings, rivers, castles, and mountains, as possible from the provided prompts in order to generate images that meet the user's expectations.

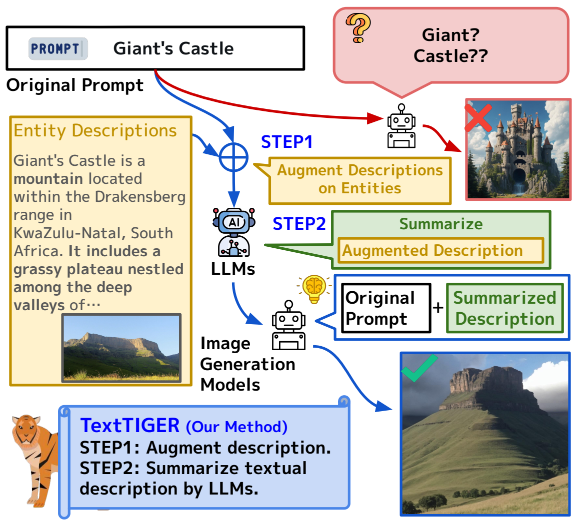


Figure 1: We propose a method, TextTIGER, which first augments descriptions of entities included in prompts and then adjusts their sequence length properly through summarization with LLMs for generating images.

However, even massive image generation models struggle to retain this knowledge or continuously acquire the latest information fully (Martinelli et al., 2024). Properly understanding entities in models helps generate user-desired images in tasks such as advertisement image generation (Mita et al., 2023). To completely incorporate up-to-date knowledge, one would need to invest substantial costs in continuously collecting data and retraining the image generation models, which is not realistic or almost impossible. For example, as shown in Figure 1, when given the prompt "`Giant's Castle`," the image generation model fails to properly understand the entity[1], i.e., "Giant's Castle (See: `https://en.wikipedia.org/`

[1] We define entity as the named entity level, which is not abstract concepts like "bridge", but specific instances such as "Golden Gate Bridge." (Choi et al., 2018; Pakhale, 2023)

wiki/Giant%27s_Castle).” Moreover, simply appending externally acquired information as a long-context prompt does not allow the Transformer (Vaswani et al., 2017) architecture to handle the information effectively and correctly (Beltagy et al., 2020; Bertsch et al., 2023) due to its maximum token length, e.g., 512 tokens.

To address the challenges posed by insufficient entity understanding in image generation, we first construct a new dataset, *WiT-Cub* (WiT with Captions and Uncomplicated Background-explanations) for the validation. WiT-Cub consists of image-caption pairs annotated with entity mentions and enriched with informative descriptions, enabling systematic evaluation of how external knowledge about entities affects quality.

Building on WiT-Cub, we propose a novel method called Text-based Intelligent Generation with Entity prompt Refinement, or TEXTTIGER. Our approach begins by retrieving entity-specific knowledge from external sources to augment the original prompt. For instance, as shown in Figure 1, for the prompt “`Giant's Castle`,” we obtain additional context such as “Giant’s Castle is a mountain located within the...” to overcome limitations in the model’s internal knowledge. We then leverage Large Language Models (LLMs) (Abdin et al., 2024; Guo et al., 2025; Team et al., 2024) to summarize these descriptions concisely, ensuring that essential information is preserved while keeping the prompt within a manageable token length. This refined prompt is then used to generate images, effectively mitigating both the model’s knowledge limitations and its difficulty in processing long contexts.

Experimental results using multiple different image generation models and LLMs on the WiT-Cub show that our method significantly outperforms baselines in widely used metrics, IS (Salimans et al., 2016), FID (Heusel et al., 2017), and CLIP-Score (Hessel et al., 2021). Furthermore, the results indicate a drop in performance when prompts are simply augmented by descriptions, while the performance improves when descriptions are summarized. Moreover, human evaluations confirm that the fully summarized descriptions are appropriately shortened to the appropriate length and outperform the baselines across criteria, i.e., informativeness, conciseness, and fluency. These findings not only prove that generating prompts of proper length with summarized descriptions of entities by LLMs significantly enhances image generation capabilities but also demonstrate that this approach is effective in overcoming the knowledge limitations of image generation models.

2 Related Work

2.1 Vision and Entity Knowledge

In Vision and Language (V&L) fields, the challenge of understanding visual and/or textual information often unveils V&L models’ limited generalization abilities in text generation from images for, e.g., newspapers (Lu et al., 2018; Liu et al., 2021), e-commerce (Ma et al., 2022), fashion (Rostamzadeh et al., 2018), and artworks (Bai et al., 2021; Hayashi et al., 2024; Ozaki et al., 2024). Likewise, Kamigaito et al. (2023) uncovers the lack of entity knowledge of a V&L model OFA (Wang et al., 2022) in the image generation tasks. An extensive study by Huang et al. (2024) introduced the “Kitten” benchmark to evaluate knowledge-intensive generation, leading to a finding that even the most advanced models frequently fail to generate entities with accurate visual details. In their experiments across domains like landmarks, plants, and animals, models like Stable Diffusion (Esser et al., 2024), DALL-E 3, and others produced images with large inaccuracies or missing critical features when asked to depict many real-world entities. This shortfall indicates that current diffusion models are limited by what they “know” from training data, and they lack a robust factual grounding of many specific entities.

2.2 Refinement of Prompts

Existing methods, such as those proposed by Hao et al. (2024); Zhan et al. (2024), primarily explore appropriate prompts for the improvement. While these prior works highlight the variability in appropriate prompts across models, they do not consider entity-specific and up-to-date knowledge not covered by image generation models. Hao et al. (2024) also introduced a reinforcement learning-based framework that rewrites user prompts into model-preferred ones, improving both aesthetics and alignment. Similarly, Zhan et al. (2024) formulated prompt refinement as a translation problem between user language and model language, leveraging image embeddings to pivot toward prompts that better reflect the model’s preferred input distribution. Other efforts, such as the dynamic prompt weighting mechanism by Mo et al. (2024), adapt the importance of each token and its diffusion

Caption	Entity List		Image
	Entity	Description	
Former seat of the Constitutional Court at Lord Rattanathibet's Mansion on Phahurat Road.	Phahurat Road	Phahurat or Pahurat sometimes described as Thailand's Little India, is an ethnic neighborhood surrounding Phahurat Road in Wang Burapha Phirom Subdistrict, Phra Nakhon District, Bangkok.	
	Constitutional Court	A constitutional court is a high court that deals primarily with constitutional law. Its main authority is to rule on whether laws that are challenged are in fact unconstitutional...	

Table 1: An example of our constructed dataset, *WiT-Cub*. We augment the entities included in image captions using external resources. Section 3 describes the detailed information, and Appendix E.5 provides another example.

time step to control the generation process more precisely. Mañas et al. (2024) proposed using LLMs to iteratively rewrite prompts based on feedback from previous generations, optimizing for semantic-image consistency. While these methods largely improve image quality and alignment, they primarily focus on stylistic, structural, or distributional refinement of prompts. They often operate within the model's inherent knowledge and do not explicitly address situations where factual or up-to-date entity knowledge is missing.

3 Dataset Creation: WiT-Cub

For the sake of systematically investigating whether augmenting named entities with rich descriptions improves the quality, we construct a new dataset, WiT with Captions and Background-explanations (*WiT-Cub*). While existing datasets such as WiT (Srinivasan et al., 2021) provide a large collection of image-captions pairs, they lack explicit entity-level information, limiting their usefulness in settings where understanding and visually grounding specific named entities is crucial. In real-world applications, prompts often contain proper nouns or named entities that assume background knowledge not explicitly provided in the caption. Without access to such knowledge, even advanced image generation models may hallucinate incorrect visual content, fail to capture distinctive features, or conflate similarly named entities.

To address this need, we extend the original WiT dataset by augmenting each image-caption pair with background descriptions of all named entities, retrieved via the Wikipedia API[2]. Specifically, WiT's metadata includes hyperlinks to the Wikipedia pages corresponding to entities mentioned in the captions. We programmatically follow these URLs and extract the introductory abstract of each page, which typically contains a concise yet informative summary of the entity, i.e., often covering its definition, category, origin, or salient characteristics. These abstracts serve as natural and reliable sources of contextual knowledge, especially for entities that are uncommon, ambiguous, or culturally specific. For instance, given a caption that simply states "Statue of Liberty at sunset," the Wikipedia abstract can provide clarifying information, e.g., its location, height, width, visual appearance, or symbolic significance, i.e., knowledge that is often critical for faithful image generation. To ensure consistency and quality, we filter for English-language entries and retain only the examples where both the image and the linked Wikipedia page remain accessible at the time of dataset construction. From the initial WiT corpus, we extract 2,500 valid instances that meet these criteria. Each instance in our dataset thus consists of a triplet: the original image, its corresponding caption, and the retrieved entity description. The resulting dataset, WiT-Cub, supports controlled experimentation on how access to entity-specific background knowledge affects the behavior of text-to-image generation models. Table 1 and Appendix E.5 present examples, and Appendix C.3 provides summary statistics of created dataset.

4 Proposed Method: TEXTTIGER

We propose a method that augments entity-specific knowledge for entities included in prompts using their precisely explained descriptions and then summarizes the descriptions to an appropriate length using LLMs, as shown in Figure 1. This approach

[2]https://www.mediawiki.org/wiki/API:Main_page

Method	Prompt for Image Generation
CAP-ONLY	The caption in WiT-Cub.
CAP-AUG-ONLY	The caption + Augmented knowledge from Wikipedia.
TEXTTIGER w/o LEN	The caption + Summarized description generated by LLMs.
TEXTTIGER	The caption + Summarized description generated by LLMs with the explicit token length.
ITERATIVE-TEXTTIGER	The caption + Iteratively applying TEXTTIGER (n = 3)

Table 2: Our proposed methods alongside the baseline.

effectively mitigates both the knowledge limitations of the image generation model and its serious weakness in handling long contexts. Our proposed method mainly comprises the following two steps: augmenting entities with informative descriptions and summarizing the descriptions by LLMs.

4.1 STEP 1: Augment Entities with Informative Descriptions

To ensure that the image generation model accurately understands entities, we augment entity-specific knowledge for entities in the caption using external and informative descriptions. Specifically, we extract entities in the caption using an entity list found in WiT-Cub and retrieve their description to mitigate the limitation of the model's knowledge.

4.2 STEP 2: Summarize the descriptions by LLMs

We let LLMs summarize the augmented entity-specific description from STEP 1 while retaining detailed entity information and ensuring an appropriate length. Following previous work (Juseon-Do et al., 2024), which demonstrated that explicitly specifying both input length and output token count helps LLMs manage length constraints, we adopt a similar approach for summarization. Specifically, we tokenize the augmented description from STEP 1 using CLIP (Radford et al., 2021)[3], the tokenizer of the text encoder commonly used in image generation models, and explicitly provide the token count to the LLMs. Since image generation models primarily use not only CLIP but also T5 (Raffel et al., 2020) as the text encoder, we set the output token limit to 180[4], ensuring compatibility with T5's token capacity. Appendix A.2 provides details about the token counts and the rationale for setting the limit to 180 tokens for image generation.

After applying these steps, we concatenate the summarized entity-specific description to the end of the caption, i.e., (caption + summarized description), forming a new prompt for image generation. Our preliminary experiments showed that appending the summarized entity-specific description of 180 tokens to the original caption achieved the best performance, as demonstrated in the ablation study in Appendix A.1. We refer to our proposed method as Text-based Intelligent Generation with Entity prompt Refinement, **TEXTTIGER**.

For the comparison with our proposed method, we evaluate another approach that more strictly ensures compliance with the token length limit. If the summarized description by LLMs still exceeds 180 token lengths, our work iteratively repeats STEP 2 until the length constraint is met. We define this method as **ITERATIVE-TEXTTIGER**, setting the maximum number of iterations to $n = 3$.

5 Experimental Settings

5.1 Dataset

We use the WiT-Cub in Section 3, which comprises images, captions, and entity descriptions. WiT-Cub comprises 2,500 instances, which provides a sufficiently reasonable quantity for our purpose.

5.2 Prompt Format

Prompt for Summarizing the Description We provide the prompt for letting LLMs summarize augmented entity-specific descriptions for image generation models in Appendix E.1. The summarized description begins with `SummaryStart:` and ends with `<SummaryEnd>`. We instruct the model to output these markers, and then extract the content between them using a regular expression.

Furthermore, to analyze the performance of our methods, we also try **TEXTTIGER W/O LEN**, where LLMs perform summarization without token counts being explicitly provided. This setting is likely to result in truncation due to the exceeded length of the input prompt for generation models.

Prompt for Image Generation **CAP-ONLY** uses only the original caption in WiT-Cub. **CAP-AUG-ONLY** involves extracting entities from the caption, obtaining their description from the entity list, and appending the description as a bullet-point list to the caption. The prompt of this method tends

[3] https://huggingface.co/openai/clip-vit-large-patch14

[4] We choose the default model. https://huggingface.co/stabilityai/stable-diffusion-3.5-large

Method	Description Generation	Image Generation	Encoder	IS (↑)	FID (↓)	CLIPScore (↑) Txt-Img	CLIPScore (↑) Img-Img
CAP-ONLY (Baseline)	–	dreamlike-photoreal-2.0	T5	20.57	43.29	29.94	67.91
	–	IF-I-L-v1.0	CLIP	21.66	**35.83**	30.31	67.84
	–	FLUX.1-dev	Both	23.03	43.27	29.26	66.95
	–	stable-diffusion-3.5-large		**24.03**	39.17	**31.32**	**69.96**
CAP-AUG-ONLY (Baseline)	–	dreamlike-photoreal-2.0	T5	20.93	42.88	29.58	68.02
	–	IF-I-L-v1.0	CLIP	21.34	**36.25**	30.52	68.38
	–	FLUX.1-dev	Both	22.40	42.80	29.17	67.71
	–	stable-diffusion-3.5-large		**23.87**	39.75	**30.52**	**69.34**

Table 3: Experimental results for the baselines CAP-ONLY and CAP-AUG-ONLY, which incorporates entity-specific descriptions without summarization. The red values indicate improvement compared to the baseline (CAP-ONLY) and **the bold values** highlight the best results among models. In CAP-AUG-ONLY, due to the excessive token length and subsequent truncation, the overall accuracy deteriorates, describing the importance of prompt refinement.

to become longer, leading to a truncated input to the text encoder of image generation models.

For the other three methods (i.e., TEXTTIGER methods), the prompt is formed by concatenating the caption and description. This approach is based on preliminary experiments in Appendix A.1, where inputting the concatenation of the caption and description as the prompt yields superior performance compared to using the description only. Table 2 provides all five methods, and Appendix E.2 describes the more detailed prompts.

5.3 Models

Summarization Models To summarize the augmented entity-specific description for the image generation process, we adopt the following LLMs: Llama3.1 (8B-Instruct and 70B-Instruct) (Dubey et al., 2024), Llama3.3 (70B-Instruct) (Dubey et al., 2024), and Qwen2.5 (72B-Instruct) (Yang et al., 2024). The 70B-class models (Llama and Qwen) are applied with quantization to 4-bit precision. As for TEXTTIGER W/O LEN, we also analyze using GPT-4o-mini (`gpt-4o-mini-2024-07-18`) (Achiam et al., 2023), assuming that GPT-4o-mini generates the summarized description of the appropriate length without explicit token count information. This choice adopts different model types (Qwen and Llama), varying model sizes (8B and 70B), and a proprietary model (GPT-4o-mini). Appendix B provides more details about model settings.

Image Generation Models The image generation models include: IF-I-L v1.0 (DeepFloyd, 2023), Dreamlike-photoreal-2.0 (Art, 2023), Stable Diffusion 3.5-large (Esser et al., 2024), and FLUX.1-dev (Labs, 2024) as shown in Appendix B. We chose the models based on prior research (Chen, 2023), which identified high-performing models. Besides this, our choice is also based on the idea of varying text encoders: T5 (Raffel et al., 2020) only (IF-I-L), CLIP only (Dreamlike), and a combination of both (Stable Diffusion, FLUX).

5.4 Evaluation Metrics for Image Generation

We evaluate the effectiveness of our method using widely used evaluation metrics in image generation fields, i.e., Inception Score (Salimans et al., 2016), Fréchet Inception Distance (Heusel et al., 2017), and CLIPScore (Hessel et al., 2021). Appendix C.5 provides a detailed explanation of these evaluation metrics, including notations.

Inception Score (IS) (Salimans et al., 2016) evaluates the diversity and semantic meaningfulness of generated images. It quantifies how confidently a classifier can predict labels for the generated images, while also measuring the diversity of label predictions. A higher score indicates that the generated images are both of high quality and varied.

Fréchet Inception Distance (FID) (Heusel et al., 2017) evaluates the difference between the feature distributions of generated and reference images. It extracts image features using Inception v3 (Szegedy et al., 2015b), and then measures how closely the distributions of real and generated images align. A lower FID value indicates that the generated images resemble the reference images more closely in terms of quality and realism.

CLIPScore (Img-Txt) (Hessel et al., 2021) measures the alignment between a generated image and its corresponding textual description. It computes how similar the text and image representations are by using a model trained on both modalities. A higher score means that the generated image is more semantically relevant to the given text.

Method	Description Generation	Image Generation	Encoder	IS (↑)	FID (↓)	CLIPScore (↑)	
						Txt-Img	Img-Img
TEXTTIGER (Ours)	Llama-3.1-8B-Instruct	dreamlike-photoreal-2.0	T5	21.46	42.34	30.83	68.51
		IF-I-L-v1.0	CLIP	21.27	**35.49**	30.81	68.88
		FLUX.1-dev	Both	23.49	41.92	29.87	68.56
		stable-diffusion-3.5-large		**24.11**	39.13	**32.02**	**70.02**
	Llama-3.3-70B-Instruct	dreamlike-photoreal-2.0	T5	21.20	42.20	29.94	68.44
		IF-I-L-v1.0	CLIP	22.21	**35.76**	30.68	69.05
		FLUX.1-dev	Both	23.74	42.88	29.63	68.47
		stable-diffusion-3.5-large		**24.45**	39.48	**31.79**	**70.72**
	Qwen2.5-72B-Instruct	dreamlike-photoreal-2.0	T5	21.60	42.35	30.01	68.59
		IF-I-L-v1.0	CLIP	21.99	**35.40**	30.63	69.34
		FLUX.1-dev	Both	23.34	42.11	29.74	68.48
		stable-diffusion-3.5-large		**24.39**	38.30	**31.99**	**70.34**
ITERATIVE-TEXTTIGER (Ours)	Llama-3.1-8B-Instruct	dreamlike-photoreal-2.0	T5	21.36	42.34	30.83	68.51
		IF-I-L-v1.0	CLIP	21.67	**35.63**	30.84	68.93
		FLUX.1-dev	Both	23.67	41.92	29.87	68.56
		stable-diffusion-3.5-large		**24.92**	39.13	**32.02**	**70.02**
	Llama-3.3-70B-Instruct	dreamlike-photoreal-2.0	T5	21.23	42.20	29.94	68.44
		IF-I-L-v1.0	CLIP	22.25	**35.76**	30.68	69.05
		FLUX.1-dev	Both	23.58	42.45	29.63	68.40
		stable-diffusion-3.5-large		**24.51**	39.48	**31.79**	**70.72**
	Qwen2.5-72B-Instruct	dreamlike-photoreal-2.0	T5	21.68	42.37	30.01	68.60
		IF-I-L-v1.0	CLIP	22.08	**35.63**	30.64	69.41
		FLUX.1-dev	Both	23.89	42.00	29.74	68.50
		stable-diffusion-3.5-large		**24.31**	38.30	**31.99**	**70.34**

Table 4: Experimental results of our proposed method. The notations are the same as those in Table 3. The results show the improvement. Underline value indicates that the score improvement is statistically significant ($p < 0.05$).

CLIPScore (Img-Img) compares two images instead of text and image. By calculating the similarity between two feature representations, this metric determines how visually or semantically similar they are. A higher score suggests that the two images share more visual or conceptual similarities.

Significance Test To demonstrate the statistical strength of our results, we run a significance test for TEXTTIGER and ITERATIVE-TEXTTIGER. Following prior work (Kamigaito et al., 2023), we use paired-bootstrap resampling (Koehn, 2004) as detailed in Appendix B.4.

6 Results

Overall Results Tables 3 and 4 show that our methods, i.e., TEXTTIGER and ITERATIVE-TEXTTIGER, significantly outperform the baseline CAP-ONLY in almost all cases for every metric. These results indicate the importance of capturing information about entities for text-to-image generation. Compared with our methods, the performance improvements of CAP-AUG-ONLY from CAP-ONLY are limited, indicating the necessity of using concise prompts in image generation rather than lengthy prompts. Thus, it is evident that our method TEXTTIGER, which augments entity descriptions and summarizes them to the appropriate length, is effective for image generation models.

Table 8 shows the results of generated images among all methods using Llama3.3 (70B) for the original caption, "`The River Nore at Kilkenny`." It can be observed that TEXTTIGER consistently produces images that are closer to the reference image across all image generation models when compared with CAP-ONLY. For example, TEXT-

Method	Description Generation	Image Generation	Encoder	IS (↑)	FID (↓)	CLIPScore (↑) Txt-Img	CLIPScore (↑) Img-Img
TEXTTIGER W/O LEN (Baseline)	Llama-3.1 8B-Instruct	dreamlike-photoreal-2.0	T5	20.66	42.04	30.03	68.49
		IF-I-L-v1.0	CLIP	19.52	37.25	30.81	67.83
		FLUX.1-dev	Both	23.12	42.60	29.83	67.99
		stable-diffusion-3.5-large		21.99	41.36	31.07	68.94
	Llama-3.3 70B-Instruct	dreamlike-photoreal-2.0	T5	20.82	42.10	29.97	68.51
		IF-I-L-v1.0	CLIP	20.66	37.02	30.67	68.11
		FLUX.1-dev	Both	22.93	42.21	29.22	67.12
		stable-diffusion-3.5-large		21.90	40.45	30.72	68.71
	Qwen2.5 72B-Instruct	dreamlike-photoreal-2.0	T5	21.20	42.35	29.90	68.64
		IF-I-L-v1.0	CLIP	20.31	35.88	30.58	68.61
		FLUX.1-dev	Both	23.25	41.93	29.76	68.27
		stable-diffusion-3.5-large		23.18	39.26	30.95	69.53

Table 5: Experimental results for TEXTTIGER W/O LEN, using prompts without explicit length control. The notations are the same as those in Table 3. It is evident that token truncation leads to performance degradation.

Method	Avg. # of Tokens	Num. of Violation
CAP-ONLY	26.48	0
CAP-AUG-ONLY	487.34	1,429
TEXTTIGER W/O LEN	314.15	2,117
TEXTTIGER (Ours)	118.89	0
ITERATIVE-TEXTTIGER	118.89	0

Table 6: Avg. # of token and # instances over T5 limit.

Method	Model	Perspective Informative	Perspective Concise	Perspective Fluent
CAP-ONLY	–	3.68	**3.81**	3.7
TEXT TIGER	Llama3.1 (8B)	3.71	3.38	**3.73**
	Llama3.1 (70B)	**3.82**	3.3	3.7
	Llama3.3 (70B)	3.78	3.24	3.63
	Qwen2.5 (72B)	3.76	3.35	3.64

Table 7: The average scores for human evaluation.

TIGER features a wide river at the center with buildings or houses on both sides. While the baseline can recognize the general layout, Dreamlike tends to produce images that evoke a river in the middle of a forest, suggesting that CAP-ONLY does not adequately capture the entities in the caption. In contrast, our proposed method, which augments the entity-related knowledge and summarizes it to an appropriate length, leads to images that more closely resemble the reference image.

TEXTTIGER v.s. ITERATIVE-TEXTTIGER Table 4 compares our method, TEXTTIGER, which generates summarized descriptions by directly specifying a target token length, with its iterative variant, ITERATIVE-TEXTTIGER, which refines the output up to three times to better satisfy the token limit. The improvements observed with ITERATIVE-TEXTTIGER suggest that both approaches yield nearly identical results, indicating that TEXTTIGER alone is sufficient to produce descriptions of appropriate length.

Importance of Length Control To reveal the importance of controlling prompt lengths, we analyze TEXTTIGER W/O LEN, which does not impose length constraints. Table 5 indicates the performance drop from TEXTTIGER, which aligns with the tendency of the generated token lengths. As shown in Table 6, this approach led to an average token sequence length of 314.15 with 2,117 violations, while CAP-AUG-ONLY had 487.34 tokens on average with 1,429 violations. These results demonstrate that exceeding the token length limit causes truncation, leading to performance degradation. In contrast to their failure, our methods control length, keeping prompts within the limit while preserving key information. This confirms length control is essential for an appropriate prompt design in image generation.

7 Analysis and Discussion

Human Evaluation To evaluate whether the descriptions summarized by LLMs include accurate and reliable information, we conducted human evaluation by multiple participants via MTurk (Crowston, 2012), following the guidelines from previous research (Fabbri et al., 2021). We show both cases, CAP-ONLY (caption only) and our method TEXTTIGER (caption + description), along with their corresponding reference image to ensure that annotators can evaluate them on an equal footing.

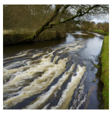
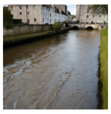
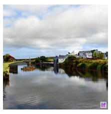
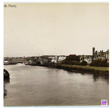
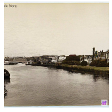
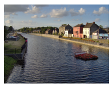
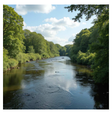
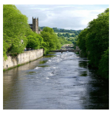
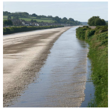
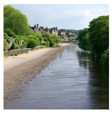
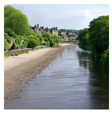
Model	CAP-ONLY	CAP-AUG-ONLY	TEXTTIGER W/O LEN	TEXTTIGER	ITERATIVE-TEXTTIGER	Reference
Dreamlike						
IF-I-L						
FLUX						
Stable Diffusion						

Table 8: The examples of outputs generated using various methods for the input "`The River Nore at Kilkenny`" alongside the reference image. The models used include Dreamlike (CLIP-only), IF-I-L (T5-only), and FLUX and Stable Diffusion, which utilize both CLIP and T5. The model used for summarization is Llama3.3 (70B).

Annotators rated them based on three criteria: Informativeness, Conciseness, and Fluency. Each criterion was scored on a scale from 1 (worst) to 5 (best), without requiring any additional explanations. Due to cost constraints, we randomly sampled 100 cases for evaluation and allocated up to 5 ($\geq$3) annotators for each case. We present the average scores for each criterion in Table 7, demonstrating that, while all models produced lower scores in conciseness compared to the baseline (CAP-ONLY), because of the description being appended, they achieved higher scores in informativeness and fluency. This suggests that the summarized descriptions by LLMs preserve more information. However, we observed only a small correlation between these human evaluation results and the performance of the image generation models, indicating that descriptions judged informative and fluent by humans do not necessarily align with improved performance in image generation models. Appendices C.2 and E.3 describe the more details.

Performance for Different Encoder Types Table 4 shows the results of image generation models using only CLIP, only T5, or both as text encoders. Comparing the Dreamlike and IF-I-L models, IF-I-L, which incorporates CLIP, consistently outperformed Dreamlike, indicating that CLIP has a greater impact on image generation than T5. However, when comparing IF-I-L with Stable Diffusion (or FLUX), models utilizing both demonstrated superior performance. This highlights the continuing importance of T5's expressive capabilities and the meaningful contribution of retaining T5 in the model effectively. From such kind of conclusions, these findings underscore the importance of maximizing information within a proper token sequence length. The results emphasize the impact of the proposed method for improving image generation.

8 Conclusion

We addressed the limitations of current text-to-image generation models in handling entity-specific knowledge, which is essential for producing accurate and user-intended outputs.

To systematically investigate this problem, we introduced *WiT-Cub*, a novel dataset that enriches image–captions pairs with entity annotations and detailed descriptions. Leveraging this dataset, we proposed TEXTTIGER, a method that augments prompts with externally retrieved entity knowledge and uses Large Language Models to summarize the information concisely, ensuring the inclusion of essential knowledge while keeping the prompt within a length suitable for image generation models.

Our experiments demonstrated that TEXTTIGER consistently outperforms baseline approaches across both automatic metrics and human evaluations, particularly in informativeness and fluency. These results confirm that entity-aware prompt refinement is a promising direction for improving factual accuracy and reliability. Our findings also highlight the potential of combining external knowledge sources with LLM-based summarization to overcome knowledge limitations.

9 Limitations

Evaluation of Object Recognition As discussed in Appendix A.6, our study evaluates the proposed method using standard evaluation metrics. These metrics primarily assess the overall diversity of generated images and the similarity of their distribution to the target distribution, e.g., via KL divergence. However, they do not directly evaluate object-level recognition within individual images. Evaluating entity-level object recognition, such as recognizing complex entities described in WiT-Cub or WiT captions, requires new evaluation metrics. Current metrics for such evaluation remain limited, and developing them represents an opportunity for future research. Our study focuses on improving image generation capabilities, leaving metric development outside our scope. On the other hand, as shown in Appendix 2, models still fail to correctly handle entities such as proper names of people, character names, and specific company names.

Limitations of Human Evaluation and Annotator Bias Annotators may have sufficient knowledge about their own country or culture but often lack familiarity with entities from other regions, leaving potential bias, especially in tasks requiring recognition of named entities from diverse geographical and cultural contexts. As future work, recruiting local annotators for each region could address this issue more effectively by ensuring that evaluators have the necessary knowledge. Furthermore, we intentionally avoided human evaluation of generated images for the following reasons. First, it is difficult to find annotators who can accurately judge entities from around the world. Second, when annotators oversimplify their judgments to reduce effort, the reliability of the evaluation deteriorates. Third, evaluating how well the generated images reflect the entities described in WiT-Cub captions demands a deep understanding of those entities. For example, a Chinese evaluator is unlikely to recognize the names of rivers, castles, or mountains in a remote region of the United States (Mostafazadeh Davani et al., 2024; Lee et al., 2024). Due to these issues, we deliberately opted not to perform human evaluations and leave it as our future studies.

Differences from Prior Work Previous studies have proposed several methods to enhance image generation capabilities. However, many of them pursue different goals and thus diverge from our approach. Lyu et al. (2024) improved image generation by leveraging multiple modalities, including speech, to infer and generate complex visual outputs. Jeong et al. (2025) improved image generation for cultural nouns through multiple refinement steps, rather than focusing on entities. Chen et al. (2022) enhanced abstract image generation via multimodal retrieval, without targeting specific entities. None of these studies deal with concrete entities at the level of specificity that we target, which makes our approach distinct.

NER for Prompt to Extract Entities We used an API to extract entities from captions and augmented them. By utilizing techniques such as Named Entity Recognition (NER) (Pakhale, 2023) to extract entities, we believe it is possible to apply this approach to a wider range of tasks (Yamada et al., 2020; Lample et al., 2016). Our focus is on enhancing image generation capabilities by expanding entity information using Wikipedia. Thus, evaluating NER itself is beyond the scope of our study, and we do not conduct such an evaluation. Additionally, we have created WiT-Cub dataset.

Comparison with Retrieval-base Methods Our method may be comparable to Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). However, our task specifically focuses on whether the performance of image generation models improves, rather than evaluating the correctness of retrieved information or competing on retrieval quality. Thus, such comparisons fall outside the scope of our work, and employing a suitable RAG system remains a promising direction for future work.

10 Ethical Considerations

When conducting human evaluation, we ensure that all 100 sampled images can be assessed fairly and that none of them violate human rights. Although MTurk[5] allows specifying the worker's race when outsourcing tasks, it is impossible to guarantee that the specified individual is the one actually performing the task (Karpinska et al., 2021; Tang et al., 2022; Gilardi et al., 2023). However, as previously mentioned, we carefully verified the 100 sampled images, making it unlikely that annotators intentionally lowered the rankings. Additionally, 3–5 individuals participate in the evaluation, ensuring the reliability of the results.

[5] https://www.mturk.com/

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, et al. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

Dreamlike Art. 2023. Dreamlike photoreal 2.0. Available at https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0.

Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5422–5432.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *Preprint*, arXiv:2305.01625.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Yixiong Chen. 2023. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings*, pages 210–221. Springer.

DeepFloyd. 2023. If-i-xl-v1.0. Available at https://huggingface.co/DeepFloyd/IF-I-XL-v1.0.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models.

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

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Preprint*, arXiv:2007.12626.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2024. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Kazuki Hayashi, Kazuma Onishi, Toma Suzuki, Yusuke Ide, Seiji Gobara, Shigeki Saito, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. IRR: Image review ranking framework for evaluating vision-language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9939–9956, Abu Dhabi, UAE. Association for Computational Linguistics.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Hsin-Ping Huang, Xinyi Wang, Yonatan Bitton, Hagai Taitelbaum, Gaurav Singh Tomar, Ming-Wei Chang, Xuhui Jia, Kelvin CK Chan, Hexiang Hu, Yu-Chuan Su, et al. 2024. Kitten: A knowledge-intensive evaluation of image generation on visual entities. *arXiv preprint arXiv:2410.11824*.

Suchae Jeong, Inseong Choi, Youngsik Yun, and Jihie Kim. 2025. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9543–9573, Albuquerque, New Mexico. Association for Computational Linguistics.

Juseon-Do, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2024. InstructCMP: Length control in sentence compression through instruction-based large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8980–8996, Bangkok, Thailand. Association for Computational Linguistics.

Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2023. Table and image generation for investigating knowledge of entities in pre-trained vision and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1904–1917, Toronto, Canada. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. .

Black Forest Labs. 2024. Flux.1-dev. Available at `https://huggingface.co/black-forest-labs/FLUX.1-dev`.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware image caption generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4013–4023, Brussels, Belgium. Association for Computational Linguistics.

Yuanhuiyi Lyu, Xu Zheng, and Lin Wang. 2024. Image anything: Towards reasoning-coherent and training-free multi-modal image generation. *arXiv preprint arXiv:2401.17664*.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18051–18061.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. 2024. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*.

Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. 2024. CNER: Concept and named entity recognition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8336–8351, Mexico City, Mexico. Association for Computational Linguistics.

Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. 2023. Camera: A multimodal dataset and benchmark for ad text generation. *arXiv preprint arXiv:2309.12030*.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. 2024. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26627–26636.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards cross-lingual explanation of artwork in large-scale vision language models. *Preprint*, arXiv:2409.01584.

Kalyani Pakhale. 2023. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv preprint arXiv:2309.14084*.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *Preprint*, arXiv:1806.08317.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

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

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015a. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015b. Rethinking the inception architecture for computer vision. corr abs/1512.00567 (2015).

Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Eighteenth symposium on usable privacy and security (SOUPS 2022)*, pages 367–385.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yogarshi Vyas and Miguel Ballesteros. 2021. Linking entities to unseen knowledge bases with arbitrary schemas. In *Proceedings of the 2021 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies*, pages 834–844, Online. Association for Computational Linguistics.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report.

Jingtao Zhan, Qingyao Ai, Yiqun Liu, Yingwei Pan, Ting Yao, Jiaxin Mao, Shaoping Ma, and Tao Mei. 2024. Prompt refinement with image pivot for text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 941–954, Bangkok, Thailand. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Ablation Study

In our preliminary experiments, we attempted to generate appropriate summarized for image generation by including captions. Table 10 presents the results, showing that prompts for image generation without including the caption led to a decline in image generation performance. This finding highlights the large impact of the 77-token limit processed by CLIP. Based on this preliminary experiment, we propose a method that supplements captions without altering them, i.e., (caption + summarized descriptions), as shown in Appendix E.2.

A.2 Why Was the Token Limit Set to 180?

As discussed in Appendix A.1, our preliminary experiment confirmed that concatenating augmented entity-specific descriptions with the original caption, i.e., (caption + description), improves performance as prompts for image generation. In our study, we limit the summary length to 180 tokens, taking the caption length, which has dozens of tokens into account. Specifically, this value is determined by subtracting the length of the caption from the maximum token limit of 256 accepted by T5. This constraint ensures that both the caption and the augmented information are fully included, enabling effective image generation.

A.3 The result of Llama3.1 (70B)

The experimental results using Llama3.1 (70B) are shown in Table 9. Based on the results in Table 4 and Table 5, it is emphasized that our method, i.e., summarization to an appropriate length using LLMs, is effective regardless of the number of model parameters when compared to Llama3.1 (8B). At the same time, it is confirmed that performance declines when the summary becomes excessively long.

A.4 Why Did GPT-4o Perform Worse?

Table 9 also shows that the result generated by gpt-4o was bad. One clear issue was that the model failed to respect the text token limit we had set. Although we specified a maximum number of new tokens, truncation still occurred mid-sentence. As a result, the image generation model received incomplete inputs, which likely led to a failure in properly understanding the prompt, i.e., this was the most critical factor affecting performance.

A.5 Token Limit

Table 6 shows the tokenized lengths of image generation prompts in each dataset, computed using the T5 tokenizer. Our method converts prompts to appropriate lengths so that they do not exceed

Method	Description Generation	Image Generation	Encoder	IS (↑)	FID (↓)	CLIPScore (↑) Txt-Img	CLIPScore (↑) Img-Img
TEXTTIGER (Ours)	Llama-3.1 70B-Instruct	dreamlike-photoreal-2.0	T5	21.08	42.10	30.81	68.67
		IF-I-L-v1.0	CLIP	22.53	**35.60**	30.66	68.88
		FLUX.1-dev	Both	23.85	42.39	29.80	68.80
		stable-diffusion-3.5-large	Both	**24.92**	39.07	**31.86**	**70.23**
ITERATOIVE-TEXTTIGER (Ours)	Llama-3.1 70B-Instruct	dreamlike-photoreal-2.0	T5	21.04	42.10	30.81	68.67
		IF-I-L-v1.0	CLIP	21.76	**35.60**	30.66	69.54
		FLUX.1-dev	Both	23.98	42.25	29.79	68.87
		stable-diffusion-3.5-large	Both	**24.03**	39.07	**31.86**	**70.23**
TEXTTIGER W/O LEN (Baseline)	Llama-3.1 70B-Instruct	dreamlike-photoreal-2.0	T5	21.63	42.61	29.95	68.36
		IF-I-L-v1.0	CLIP	21.13	36.08	30.67	69.02
		FLUX.1-dev	Both	22.85	42.51	29.85	68.37
		stable-diffusion-3.5-large	Both	23.79	39.17	31.09	69.90
	GPT-4o mini	dreamlike-photoreal-2.0	T5	18.41	47.13	26.55	62.89
		IF-I-L-v1.0	CLIP	19.09	41.70	26.75	61.73
		FLUX.1-dev	Both	16.96	59.75	26.17	59.41
		stable-diffusion-3.5-large	Both	14.99	68.10	27.09	60.39

Table 9: The experimental results obtained using Llama 3.1 (70B) and GPT-4o-mini.

the maximum sequence length supported by the T5-based image generation model.

A.6 Object Recognition

Table 8 and Appendix E.7 show the images generated by image generation models. While some images deviate from the reference images, others bear a strong resemblance.

A.7 Generalization to Unseen Entities

While TEXTTIGER improves image generation by augmenting and summarizing entity-specific knowledge, its effectiveness depends on the availability and quality of external knowledge sources, such as Wikipedia. When encountering entities with limited or no publicly available descriptions, the method may struggle to provide meaningful augmentations, potentially reducing its advantage over baseline methods (Vyas and Ballesteros, 2021; Zhang et al., 2022; Logeswaran et al., 2019).

A.8 Do LLMs Generate Summarized Descriptions Correctly?

To generate summarized descriptions for image generation, we instructed the model to output start and end markers, i.e., `SummaryStart:` and `<SummaryEnd>` as shown in Appendix E.1, and used only the text extracted between them for image generation. This approach enabled the creation of descriptions that were both of appropriate length and properly summarized for image generation. Appendix E.6 provides examples of the summarized descriptions output by the model and the corresponding images generated using them.

B Detailed Model Settings

B.1 LLMs

The table below provides detailed configurations of the models used in this study. For LLM inference to create properly summarized descriptions, we set the seed to 0. The max_tokens varied by method: 512 tokens for TEXTTIGER W/O LEN and 180 tokens for TEXTTIGER and ITERATIVE-TEXTTIGER. During image generation, we fixed the seed at 42. We conducted the experiments using Transformers library (Wolf et al., 2020) and applied quantization with bitsandbytes[6]. For OpenAI API usage, we processed requests in batches, setting max_tokens to 512 and the seed to 0. Processing all TEXTTIGER W/O LEN experiments costs approximately $10.

B.2 Image Generation Models

For image generation, we followed the configuration of Stable Diffusion 3.5. The model generates

[6]https://github.com/bitsandbytes-foundation/bitsandbytes

Method	Description Generation	Image Generation	Encoder	IS (↑)	FID (↓)	CLIPScore (↑) Txt-Img	Img-Img
CAP-ONLY	–	Dreamlike	T5	20.57	43.29	29.94	67.91
	–	IF-I-L	CLIP	21.66	35.83	30.31	67.84
	–	Stable Diffusion	Both	24.03	39.17	31.32	69.96
TEXTTIGER	Llama-3.1 8B-Instruct	Dreamlike	T5	19.75	48.51	29.93	68.51
		IF-I-L	CLIP	21.95	38.91	30.81	68.88
		Stable Diffusion	Both	22.14	43.11	31.12	70.02
	Llama-3.3 70B-Instruct	Dreamlike	T5	19.51	45.67	29.94	68.44
		IF-I-L	CLIP	22.10	37.66	30.68	69.05
		Stable Diffusion	Both	23.67	41.50	30.89	69.82
	Qwen2.5 72B-Instruct	Dreamlike	T5	18.74	91.86	30.01	68.59
		IF-I-L	CLIP	16.37	59.14	30.63	69.34
		Stable Diffusion	Both	18.20	82.99	31.09	70.34
ITERATIVE-TEXTTIGER	Llama-3.1 8B-Instruct	Dreamlike	T5	19.73	48.51	29.93	68.51
		IF-I-L	CLIP	21.80	38.91	30.84	68.93
		Stable Diffusion	Both	22.01	43.11	31.12	70.02
	Llama-3.3 70B-Instruct	Dreamlike	T5	19.51	45.67	29.94	68.44
		IF-I-L	CLIP	21.96	37.66	30.68	69.05
		Stable Diffusion	Both	23.69	41.50	30.89	69.82
	Qwen2.5 72B-Instruct	Dreamlike	T5	18.62	91.86	30.01	68.60
		IF-I-L	CLIP	16.14	59.14	30.64	69.41
		Stable Diffusion	Both	18.23	82.99	31.09	70.34

Table 10: The result of our preliminary experiment among comparisons across Dreamlike, IF-I-L, and Stable Diffusion. We confirmed that using summarized captions instead of the original ones as input for image generation models resulted in lower accuracy. Our proposed method, TEXTTIGER, described in Section 4, overcomes these challenges and demonstrates improvements over the baseline.

Model	Params	HuggingFace Name / OpenAI API
LLaMA3.1	8B	meta-llama/Llama-3.1-70B-Instruct
LLaMA3.1	70B	meta-llama/Llama-3.1-70B-Instruct
LLaMA3.3	70B	meta-llama/Llama-3.3-70B-Instruct
Qwen2.5	72B	Qwen/Qwen2.5-72B
GPT-4o-mini	–	GPT-4o-mini-2024-0718
Dreamlike	–	dreamlike-art/dreamlike-photoreal-2.0s
IF-I-L	–	DeepFloyd/IF-I-L-v1.0
FLUX.1-dev	–	black-forest-labs/FLUX.1-dev
Stable Diffusion	–	stabilityai/stable-diffusion-3.5-large
T5	4.7B	google-t5/t5-11b
CLIP	428M	openai/clip-vit-large-patch14

Table 11: Detailed name of models. As for T5, only the encoder part is used in image generation models.

images with a resolution of 1,024 × 1,024 pixels. The guidance scale is set to 3.5, and the number of inference steps is 50. The maximum sequence length for processing inputs is 512 tokens.

B.3 Experimental Environments

We used the NVIDIA RTX 6000 Ada Generation to create prompts designed for appropriate image generation. For the image generation process, we employed the NVIDIA RTX 6000 Ada Generation with Stable Diffusion and FLUX, which incorporates both T5 and CLIP. We used the NVIDIA A6000 with Dreamlike and IF-I-L.

B.4 Detailed Significance Test

Following prior work (Kamigaito et al., 2023), we conducted statistical testing using paired-bootstrap resampling (Koehn, 2004). We randomly extracted 2,000 samples with replacement from the dataset and ran the test 1,000 times.

B.5 Reproducibility of Outputs

This study relies on external resources, including the OpenAI API, external LLMs, and image generation models. Changes in the availability or performance of these resources, beyond our control, could affect reproducibility. The batch processing cost for using the OpenAI API in our research was approximately $10.

C Detailed Evaluation

C.1 Details of Human Evaluation (MTurk)

We used MTurk to evaluate the summarized descriptions for image generation generated by LLMs. We compared four different LLMs with the baseline (WiT-Cub captions) and designed the evaluation procedure following previous research (Fabbri et al., 2021). Annotators rated the prompts on three criteria: (1) Informativeness, (2) Conciseness, and (3) Fluency, using a five-point scale (1 = worst, 5 = best). Details of the evaluation procedure are provided in Appendix E.3. To ensure reliability, we hired multiple annotators, with up to five annotators per question (greater than three annotators). Due to cost constraints, we sampled 100 cases for evaluation. Additionally, to maintain consistency, we intentionally included duplicate questions, i.e., dummy ones. If an annotator provided inconsistent answers for the same question, we excluded their responses from the final analysis. Each question is distributed at a rate of 3 dollars. The sum in our work is around $200. We outsourced 100 questions, offering a reward of $3 per question, with a maximum of five annotators per question. This amount also accounts for factors such as dry runs and the exclusion of inattentive annotators. We hired workers who have an approval rate greater than 90% with at least 50 approved HITs, following the prior research. (Sakai et al., 2024)

C.2 Statistics of Human Evaluation

Table 12 presents the inter-annotator agreement values measured by Fleiss' Kappa (Cohen, 1960) and Krippendorff's Alpha (Krippendorff, 2011). To ensure the reliability of annotators, as done in prior studies (Hayashi et al., 2025; Filippova et al., 2015), we exclude annotators who consistently produce outliers and use the scores from the remaining annotators. For evaluation, we use questions after removing dummy questions inserted to assess annotator reliability. Krippendorff's evaluation scale is set to "ordinal."

Metrics	Concise	Fluency	Informative
Fleiss' Kappa	0.335	0.22	0.364
Krippendorff's Alpha	0.731	0.677	0.685

Table 12: Statistics results of human evaluation.

C.3 Detailed Dataset Statistics

The WiT-Cub dataset created in our study in Section 3 is an extension of WiT. Therefore, the image resolution and size remain unchanged from the original dataset.

Detail	Value
# of instances	2,500
#Avg. number of entities	3.02
#Avg. token length	26.48

Table 13: WiT-Cub statistics. We calculate the token sequence length by CLIP, as described in Section 4.2.

C.4 Automatic Evaluation of Summarization

In our study, we did not conduct automatic evaluations for the summarization, such as ROUGE scores (Lin, 2004), for two reasons: 1) There are no reference answers for the descriptions augmented in our study, making automatic evaluation infeasible; 2) Although an exact match-based method exists for measuring how many entities are included in the generated text (Shao et al., 2024), we augmented all entities using the Wikipedia API and summarized them with LLMs. As a result, entities are guaranteed to appear in the summaries. Due to these reasons, we did not conduct automatic evaluations for the summaries. Instead, we performed large-scale human evaluations, which are more insightful than automatic metrics. The results confirmed that the summaries are informative, demonstrating the effectiveness of our method.

C.5 Detailed Evaluation Metrics

Inception Score (IS) (Salimans et al., 2016) evaluates the diversity and semantic meaningfulness of generated images. It analyzes the label distribution of images using a classifier and computes the score based on entropy and KL divergence. A higher score indicates greater diversity and quality of the generated images.

$$\mathrm{IS} = \exp\left(\mathbb{E}_{x\sim p_g}\left[D_{\mathrm{KL}}(p(y|x)||p(y))\right]\right) \quad (1)$$

Here, x represents a generated image, p_g denotes the distribution of generated images, $p(y|x)$ is the predicted label distribution for image x, $p(y)$ is the marginal label distribution over all generated images, and D_{KL} represents the KL divergence.

Fréchet Inception Distance (FID) (Heusel et al., 2017) measures the difference in feature distributions between generated and reference images. It extracts image features using the Inception network (Szegedy et al., 2015a) and calculates the Fréchet distance between the distributions. A lower value indicates higher quality and closer resemblance of generated images to real images.

$$\text{FID} = ||\mu_r - \mu_g||_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \tag{2}$$

Here, μ_r and μ_g are the mean vectors of the feature distributions for real and generated images, respectively. Σ_r and Σ_g are the covariance matrices for the feature distributions of real and generated images, Tr denotes the trace of a matrix, and $||\cdot||_2$ represents the 2-norm.

CLIPScore (Img-Txt) (Hessel et al., 2021) evaluates the relevance between generated images and text. A higher score indicates that the image aligns well with the text content.

$$\text{CLIPScore}_{\text{Img-Txt}} = \cos(E_{\text{img}}(x), E_{\text{txt}}(t)) \tag{3}$$

Here, $E_{\text{img}}(x)$ is the CLIP embedding vector for image x, $E_{\text{txt}}(t)$ is the CLIP embedding vector for text t, and $\cos(\cdot,\cdot)$ represents cosine similarity.

CLIPScore (Img-Img) evaluates the similarity between two images by calculating the cosine similarity between their CLIP embedding vectors. A higher score indicates that the two images are semantically similar.

$$\text{CLIPScore}_{\text{Img-Img}} = \cos(E_{\text{img}}(x_1), E_{\text{img}}(x_2)) \tag{4}$$

Here, $E_{\text{img}}(x_1)$ and $E_{\text{img}}(x_2)$ are the CLIP embedding vectors for images x_1 and x_2, respectively, and $\cos(\cdot,\cdot)$ represents cosine similarity.

D Case Study on Failure Cases

We still observed failure cases during prompt summarization and image generation.

D.1 Summarized Description by LLMs

TEXTTIGER instructed the model to generate specific start and end symbols for the summary, then extracted the description between them using regular expressions. The extracted text uses as the summarized description. However, in some cases, the model generated the end symbol at an incorrect position, resulting in unnatural sentences.

Failure Case of Summarization

Budapest is the capital and most populous city of Hungary. It has a population of 1,752,286 and is the centre of the Budapest metropolitan area. The city has a rich history, dating back to the Roman town of Aquincum, and has been influenced by various cultures, including the Hungarians, Mongols, and Ottomans. Budapest became a global city after the unification of Buda, Óbuda, and Pest in 1873. The city is a hub for commerce, finance, media, art, fashion, research, technology, education, and entertainment. It is home to over 40 colleges and universities, including Eötvös Loránd University and the Budapest University of Technology and Economics. The city's central area along the Danube River is a UNESCO World Heritage Site and features several notable monuments of classical architecture. Budapest attracts around

D.2 Image generation

We observed cases where this approach was less effective, leading to images that deviated from the reference. Figure 2 shows an example where the caption "`An electronic billboard on the Thomson Reuters building welcomes Facebook to the Nasdaq.`" was used. These results indicate that while our method effectively enhances entity-based descriptions for scenes, landscapes, and buildings, it struggles to accurately reflect faces or texts. Table 13 provides the statistics of WiT-Cub.

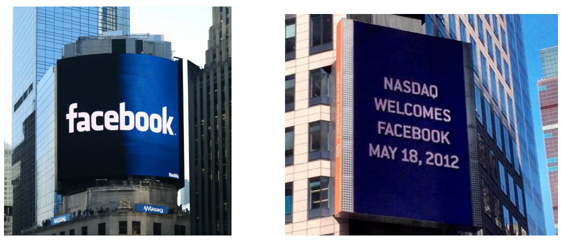


Figure 2: (Left) Generated image by Stable Diffusion and (Right) reference image.

E AI Assistant Tools

We leverage advanced tools such as ChatGPT [7], DeepL [8], and GitHub Copilot [9]to accelerate our research processes and streamline the writing of academic papers. These tools enable efficient idea generation, precise translations, and coding assistance, enhancing productivity and quality.

[7]`https://openai.com/index/chatgpt/`
[8]`https://www.deepl.com/en/translator`
[9]`https://github.com/features/copilot`

E.1 Prompt for Summarization

The prompts used as input to the LLMs for generating properly summarized descriptions for image generation are shown below, where blue text represents variables and red text indicates the explicit input of token counts tokenized by CLIP (https://huggingface.co/openai/clip-vit-large-patch14).

Prompt for Summarizing (TextTIGER w/o Len)

```
Please generate a summary so that there are 180 tokens.
However, please do not delete proper nouns or other important information.
Please begin the output with SummaryStart: and write the summary of the text.
Please end the output with <SummaryEnd> as the last token.

Example:
SummaryStart: The summary of the text is as follows. The text is about the summary of the text.
<SummaryEnd>

Complement:
{Complement}

SummaryStart:
```

Prompt for Summarizing (TextTIGER)

```
The current tokens are {current_words} tokens.
Please generate a summary so that there are 180 tokens.
However, please do not delete proper nouns or other important information.
Please begin the output with SummaryStart: and write the summary of the text.
Please end the output with <SummaryEnd> as the last token.

Example:
SummaryStart: The summary of the text is as follows. The text is about the prompt of the text.
<SummaryEnd>

Complement:
{Complement}

SummaryStart:
```

Prompt for Summarizing (Iterative-TextTIGER)

```
The current tokens are still {current_words} tokens.
Please generate a summary so that there are 180 tokens.
However, please do not delete proper nouns or other important information.
Please begin the output with SummaryStart: and write the summary of the text.
Please end the output with <SummaryEnd> as the last token.

Example:
SummaryStart: The summary of the text is as follows. The text is about the prompt of the text.
<SummaryEnd>

Complement:
{description}

SummaryStart:
```

E.2 Prompt for Image Generation

The prompts used for image generation are as follows. To maximize the information content, we only include the necessary information. Blue text represents variables.

Prompt for Image Generation (CAP-ONLY)

```
Caption: {caption}
```

Prompt for Image Generation (CAP-AUG-ONLY and Three TEXT-TIGER Methods)

```
Caption: {caption}

Note: {description}
```

E.3 Details of Human Evaluation

Below, we provide the procedure used for outsourcing evaluations via Amazon Mechanical Turk (MTurk (Crowston, 2012), https://www.mturk.com/). The procedure was designed with reference to previous research on summarization evaluation (Fabbri et al., 2021). For each task, we hired up to five evaluators on MTurk. Additionally, to ensure the reliability of their assessments, we included identical test cases within the evaluation subset to verify consistency in their responses.

Prompt for Image Generation

```
# Instructions
In this task, you will evaluate how well the provided captions match the given images.
To complete this task correctly, follow these steps:
  1. Watch the image and understand the scene.
  2. Read the caption and compare it with the image.
  3. Rate the caption based on the following criteria on a scale from 1 (worst) to 5 (best):
  4. Please only score the rank without explaining the reason.


# Definitions
## Informativeness:
  - How much useful information the caption provides about the image.
  - Captions should include relevant details, such as proper nouns and contextual information,
to help the reader visualize the image.

## Conciseness:
  - How accurately and efficiently the caption describes the image.
  - It should avoid unnecessary details while clearly conveying the key points.

## Fluency:
  - How natural and well-structured the caption is.
  - It should be a coherent sentence rather than a list of words.
```

E.4 A Sample of MTurk

The figure below indicates a sample screenshot of MTurk.

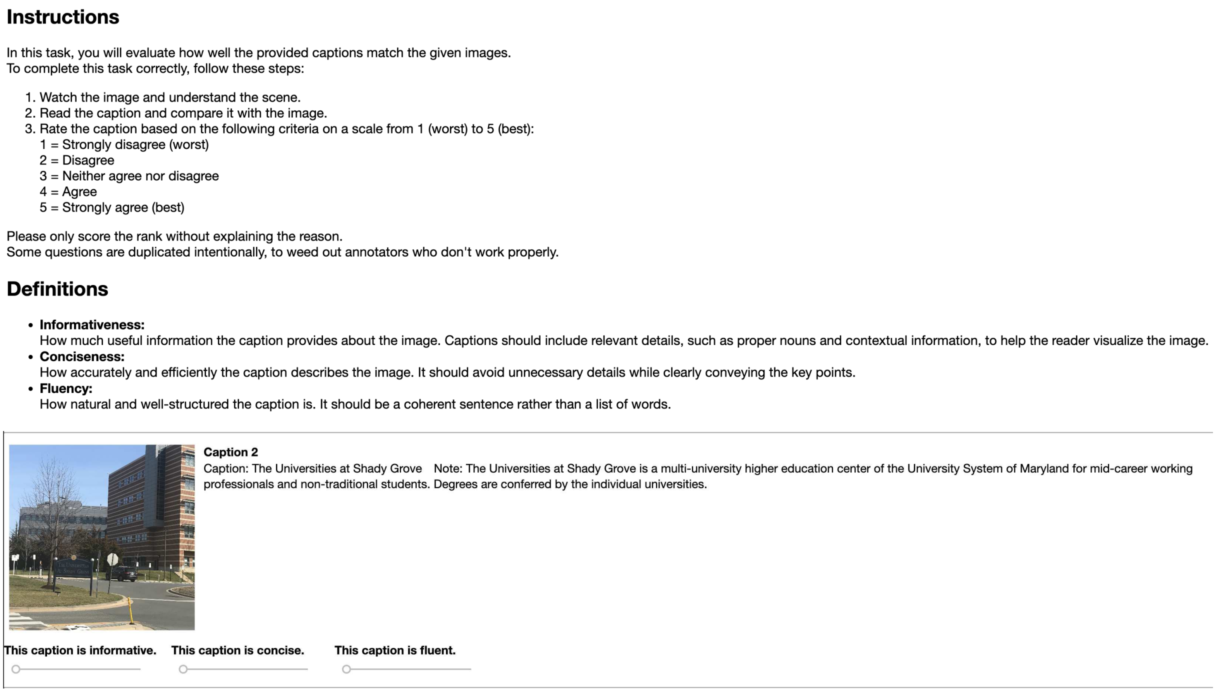


Figure 3: A sample screenshot in MTurk.

E.5 Another Example of WiT-Cub

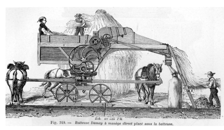
Caption	Entity List: Entity	Entity List: Description	Image
Threshing machine in 1881. Today both reaping and threshing are done with a combine harvester.	threshing machine	A threshing machine or a thresher is a piece of farm equipment that separates grain seed from the stalks and ...	
	combine harvester	The modern combine harvester, also called a combine, is a machine designed to harvest a variety of cultivated seeds. Combine harvesters are one of the most....	
	machine	A machine is a physical system ...	

Table 14: Another example of the dataset, *WiT-Cub*, we created. We augment the entities included in image captions using external resources. Section 3 demonstrate the dataset creation process.

E.6 Examples of the Summarized Description

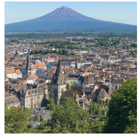
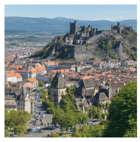
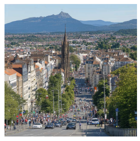
Caption	LLMs	Summarized description	Generated image
		Reference 	
Clermont Ferrand	Llama 3.1 8B-Instruct	Clermont-Ferrand is a city in France with a population of 147,284. It is the prefecture of the Puy-de-Dôme department and has a metropolitan area of 504,157 inhabitants. The city is known for its chain of volcanoes, the Chaîne des Puys, and its dormant volcano Puy de Dôme. Clermont-Ferrand has been listed as a UNESCO World Heritage Site since 2018. The city has a rich history, dating back to the Gallo-Roman era, and has been the site of several important events, including the Council of Clermont in 1095. Today, Clermont-Ferrand hosts the Clermont-Ferrand International Short Film Festival and is home to the corporate headquarters of Michelin. The city is also a learning city, with a quarter of its population being students.	
	Llama 3.1 70B-Instruct	Clermont-Ferrand is a city and commune of France, in the Auvergne-Rhône-Alpes region, with a population of 147,284 (2020). It is the prefecture of the Puy-de-Dôme département and has a metropolitan area of 504,157 inhabitants. The city is known for the chain of volcanoes, the Chaîne des Puys, and the dormant volcano Puy de Dôme. Clermont-Ferrand has been listed as a ẗectonic hotspoẗsince July 2018 on the UNESCO World Heritage List. The city has a rich history, dating back to the Gallo-Roman era, and has been an important center for trade and culture. Today, Clermont-Ferrand hosts the Clermont-Ferrand International Short Film Festival and is home to the corporate headquarters of Michelin.	
	Llama 3.3 70B-Instruct	Clermont-Ferrand is a city and commune of France, in the Auvergne-Rhône-Alpes region, with a population of 147,284. It is the prefecture of the Puy-de-Dôme département and has a metropolitan area with 504,157 inhabitants. The city is known for the chain of volcanoes, the Chaîne des Puys, and has been listed as a ẗectonic hotspoẗsince July 2018 on the UNESCO World Heritage List. Clermont-Ferrand has a rich history, dating back to the Gallo-Roman era, and has been an important city in France, hosting the Council of Clermont in 1095. Today, it is home to the corporate headquarters of Michelin and hosts the Clermont-Ferrand International Short Film Festival. The city is also known for its distinctive black lava.	
	Qwen 2.5 72B-Instruct	Clermont-Ferrand, a city and commune in the Auvergne-Rhône-Alpes region of France, has a population of 147,284 (2020) and a metropolitan area of 504,157 inhabitants (2018). As the capital of the Puy-de-Dôme département, it is known for its volcanic landscape, including the Puy de Dôme, and its designation as a UNESCO World Heritage Site. Historically significant, it was the capital of the Arvernie tribe and later named Augustonemetum during the Roman era. The city played a crucial role in the First Crusade, hosting the Council of Clermont in 1095. Today, it is home to the Michelin headquarters, the Clermont-Ferrand International Short Film Festival.	

Table 15: Images generated using our proposed method (TEXTTIGER) by Stable Diffusion 3.5. This table compares simplified and summarized descriptions across LLMs which we use for generating the prompt.

Caption	LLMs	Summarized description	Generated image
		Reference 	
Construction of the Metropolitan Railway close to King's Cross station in 1861	Llama 3.1 8B-Instruct	The Metropolitan Railway was a passenger and goods railway that served London from 1863 to 1933. It connected the main-line railway termini at Paddington, Euston, and King's Cross to the City. The first section was built beneath the New Road using cut-and-cover between Paddington and King's Cross. The line was extended from both ends and northwards via a branch from Baker Street. The most important route was northwest into the Middlesex countryside, stimulating the development of new suburbs. Electric traction was introduced in 1905, and the Met developed land for housing, promoting housing estates near the railway using the ¨Metro-land¨brand. The Met was amalgamated with the Underground Electric Railways Company of London and the capital's tramway and bus operators to form the London Passenger Transport Board in 1933.	
	Llama 3.1 70B-Instruct	The Metropolitan Railway was a passenger and goods railway that served London from 1863 to 1933. It connected the main-line railway termini at Paddington, Euston, and King's Cross to the City. The first section opened on 10 January 1863 with gas-lit wooden carriages hauled by steam locomotives. The line was extended from both ends, and northwards via a branch from Baker Street. Southern branches reached Hammersmith in 1864, Richmond in 1877, and the original completed the Inner Circle in 1884. Electric traction was introduced in 1905, and by 1907 electric multiple units operated most of the services. The Met developed land for housing and promoted housing estates near the railway using the ¨Metro-land¨brand. On 1 July 1933, the Met was amalgamated with other companies.	
	Llama 3.3 70B-Instruct	The Metropolitan Railway was a passenger and goods railway that served London from 1863 to 1933. It connected the main-line railway termini at Paddington, Euston, and King's Cross to the City. The first section was built beneath the New Road using cut-and-cover between Paddington and King's Cross. The line was soon extended from both ends, and northwards via a branch from Baker Street. The Met developed land for housing and promoted housing estates near the railway using the ¨Metro-land¨brand. On 1 July 1933, the Met was amalgamated with other companies to form the London Passenger Transport Board. Former Met tracks and stations are used by the London Underground's Metropolitan, Circle, District, Hammersmith & City, Piccadilly, Jubilee and Victoria lines, and by Chiltern Railways and Great Northern.	
	Qwen 2.5 72B-Instruct	The Metropolitan Railway, or Met, began serving London in 1863, connecting major railway termini like Paddington, Euston, and King's Cross to the City. The initial line was constructed using cut-and-cover methods and tunnels, opening to the public on 10 January 1863. Extensions followed, reaching Hammersmith in 1864, Richmond in 1877, and completing the Inner Circle in 1884. The Met played a crucial role in developing suburban areas, extending to Harrow in 1880 and Verney Junction in 1897. Electric traction was introduced in 1905, and by 1907, most services were electric. The Met also engaged in property development, promoting ¨Metro-land¨housing estates.	

Table 16: Images generated using our proposed method (TEXTTIGER) by Stable Diffusion 3.5. This table compares simplified and summarized descriptions across LLMs which we use for generating the prompt.

E.7 Another Example of Generated Images

Here, we introduce some examples of generated images.[10]

Model	CAP-ONLY (Baseline)	CAP-AUG-ONLY	TEXTTIGER W/O LEN	TEXTTIGER	Iterative-TEXTTIGER	Reference
Dreamlike						
IF-I-L						
FLUX						
Stable Diffusion						

Table 17: Another example of generated images using various methods for the input "`Haymarket roundabout, Melbourne`" alongside their reference images. The models used include Dreamlike (CLIP-only), IF-I-L (T5-only), and FLUX and Stable Diffusion which utilize both as text encoders. The model used for summarization is Qwen2.5 (72B).

Model	CAP-ONLY (Baseline)	CAP-AUG-ONLY	TEXTTIGER W/O LEN	TEXTTIGER	Iterative-TEXTTIGER	Reference
Dreamlike						
IF-I-L						
FLUX						
Stable Diffusion						

Table 18: Another example of generated images using various methods for the input "`Smelting Works. Oreana, Nevada. ca. 1857 by Timothy H. O'Sullivan.`" alongside their reference images. The models used include Dreamlike (CLIP-only), IF-I-L (T5-only), and FLUX and Stable Diffusion which utilize both as text encoders. The model used for summarization is Qwen2.5 (72B).

[10] Due to reduced resolution for file size constraints, some images may appear blurry or hard to see.

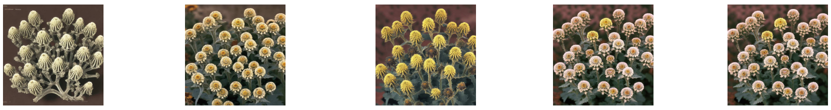
Model	CAP-ONLY (Baseline)	CAP-AUG-ONLY	TEXTTIGER W/O LEN	TEXTTIGER	Iterative-TEXTTIGER	Reference
Dreamlike						
IF-I-L						
FLUX						
Stable Diffusion						

Table 19: Another example of generated images using various methods for the input “`Helichrysum arenarium from Thomé Flora von Deutschland, Österreich und der Schweiz 1885`” alongside their reference images. The models used include Dreamlike (CLIP-only), IF-I-L (T5-only), and FLUX and Stable Diffusion which utilize both as text encoders. The model used for summarization is Qwen2.5 (72B).

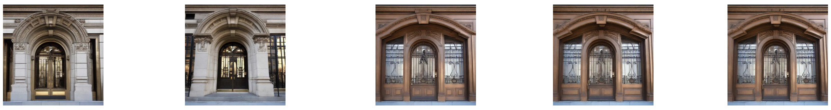
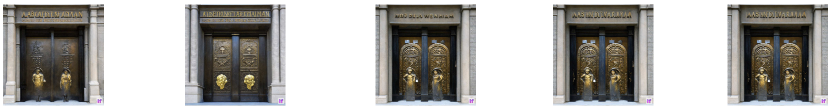
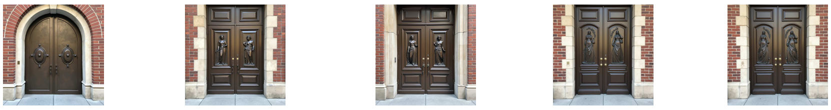
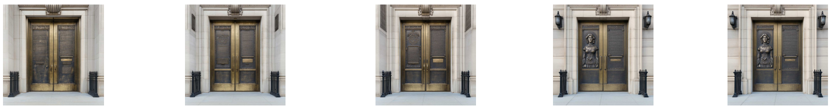
Model	CAP-ONLY (Baseline)	CAP-AUG-ONLY	TEXTTIGER W/O LEN	TEXTTIGER	Iterative-TEXTTIGER	Reference
Dreamlike						
IF-I-L						
FLUX						
Stable Diffusion						

Table 20: Another example of generated images using various methods for the input “`The bronze entrance doors to the administration building on West 155th Street were designed by Academy member Adolph Alexander Weinmaen.`” alongside their reference images. The models used include Dreamlike (CLIP-only), IF-I-L (T5-only), and FLUX and Stable Diffusion which utilize both as text encoders. The model used for summarization is Qwen2.5 (72B).