Harmonizing Visual Text Comprehension and Generation

Zhen Zhao[1,2,*], Jingqun Tang[2,✉], Binghong Wu[2], Chunhui Lin[2], Shu Wei[2], Hao Liu[2], Xin Tan[1], Zhizhong Zhang[1,3], Can Huang[2], Yuan Xie[1,✉]
[1]East China Normal University [2] ByteDance
[3] Shanghai Key Laboratory of Computer Software Evaluating and Testing, Shanghai, China
51255901056@stu.ecnu.edu.cn, tangjingqun@bytedance.com, yxie@cs.ecnu.edu.cn

Abstract

In this work, we present TextHarmony, a unified and versatile multimodal generative model proficient in comprehending and generating visual text. Simultaneously generating images and texts typically results in performance degradation due to the inherent inconsistency between vision and language modalities. To overcome this challenge, existing approaches resort to modality-specific data for supervised fine-tuning, necessitating distinct model instances. We propose Slide-LoRA, which dynamically aggregates modality-specific and modality-agnostic LoRA experts, partially decoupling the multimodal generation space. Slide-LoRA harmonizes the generation of vision and language within a singular model instance, thereby facilitating a more unified generative process. Additionally, we develop a high-quality image caption dataset, DetailedTextCaps-100K, synthesized with a sophisticated closed-source MLLM to enhance visual text generation capabilities further. Comprehensive experiments across various benchmarks demonstrate the effectiveness of the proposed approach. Empowered by Slide-LoRA, TextHarmony achieves comparable performance to modality-specific fine-tuning results with only a 2% increase in parameters and shows an average improvement of 2.5% in visual text comprehension tasks and 4.0% in visual text generation tasks. Our work delineates the viability of an integrated approach to multimodal generation within the visual text domain, setting a foundation for subsequent inquiries. Code is available at `https://github.com/bytedance/TextHarmony`.

1 Introduction

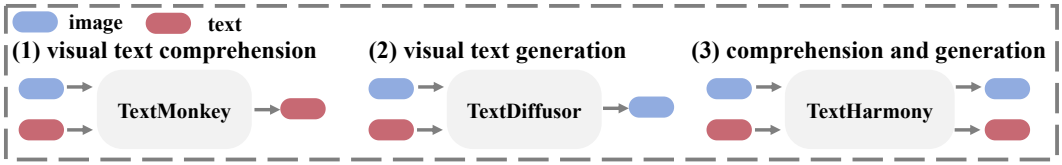
Visual text comprehension and generation tasks such as scene text detection and recognition [79, 63, 40, 75, 62, 61, 24, 58], document understanding [64, 23], visual question answering (VQA) [26, 15, 27, 39, 59], key information extraction (KIE) [64, 23], multi-modal retrieval[33, 34, 30, 31, 36, 29, 32, 35, 28], visual text generation, editing, and erasure [66, 6, 5] are consistently of significant value for both academic research and practical applications. Recently, remarkable advancements have been achieved in visual text comprehension and generation, driven by the evolution of Multimodal Large Language Models (MLLMs) and diffusion models. Foremost text-centric MLLMs [73, 20, 27, 39] utilize a cohesive framework to comprehend text-rich images comprehensively, whereas diffusion-based approaches [66, 6, 5] introduce innovative modifications to enhance visual text generation capabilities. As depicted in Figure 1, text-centric MLLMs and diffusion models are capable of handling language and vision modalities adeptly, with MLLMs generating texts and diffusion models producing images. However, integrating language and vision generation capabilities within a large multimodal model for visual text scenarios remains unexplored. This paper focuses on

[*] Work done when Zhen Zhao was an intern at ByteDance. ✉ Corresponding authors.

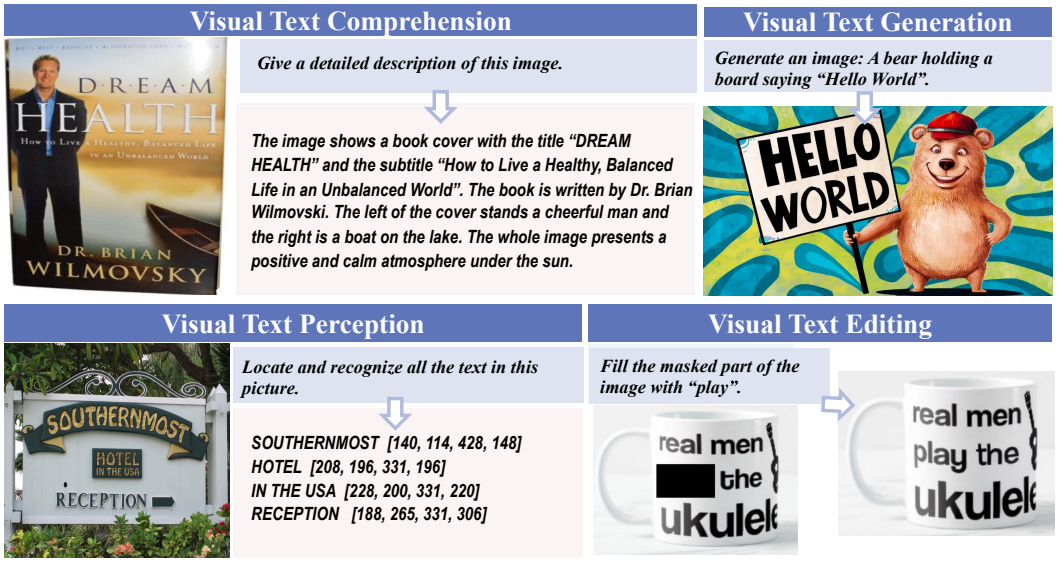
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3093

the simultaneous manipulation of language and vision generations to further streamline the processing of diverse text-centric multimodal tasks.



(a) Comparison of model pipelines



(b) Visual text-related multimodal generation tasks

Figure 1: Figure (a) illustrates the different types of image-text generation models: visual text comprehension models can only generate text, visual text generation models can only generate images, and TextHarmony can generate both text and images. Figure (b) illustrates the versatility of TextHarmony in generating different modalities for various text-centric tasks.

In the general multimodal domain, some pioneering efforts [57, 16, 80, 65] empower MLLMs with the ability to generate images beyond texts, vastly extending the versatility of multimodal models. Such advancements inspire us to develop a text-centric multimodal generative model. Our foundational model follows these approaches, incorporating a VIT-based image encoder, a text tokenizer, an LLM, a text detokenizer, and a diffusion-based image decoder.

Previous works [57, 80, 65] and our pilot experiments (Figure 2) have shown that multimodal generation often leads to a notable decline in performance due to the substantial inconsistency between language and vision modalities in the generation space. Prior studies [57, 16, 80, 65] commonly rely on modality-specific supervised fine-tuning to bolster generative capacities. A nuanced challenge involves boosting generative capabilities across modalities using a singular model instance. Mixed-of-Experts (MoE) [54] architecture is widely adopted in LLMs because it improves the model's scalability while keeping the cost of inference similar to that of smaller models. Impressed by MOE-based models [18, 9] that set up task-specific experts to handle different tasks efficiently, we propose adapting modality-specific experts to partially decouple the generation of images and texts. Transforming a dense multimodal generative framework into an MoE-based sparse model poses significant challenges due to the high computational demand and extensive training data requirements.

To tackle these challenges, we incorporate Low-Rank Adaptation(LoRA) [22] experts instead. Specifically, we integrate multiple LoRA experts into the vision encoder and LLM components, encompassing modality-agnostic experts, vision modality generation experts, and language modality generation experts. Modality-specific experts are instrumental in refining and integrating modality-specific generative representations, while modality-agnostic experts enhance certain general representations. A dynamic gating network then amalgamates these modality-specific and general generative rep-

resentations, assigning precise expert weights. Consequently, with minimal parameter increase, we enhance image comprehension and generation capabilities within a singular model instance. Slide-LoRA achieves results comparable to those obtained through separate modality-specific fine-tuning, demonstrating the efficacy of our approach in bridging the gap between language and vision modalities in multimodal generation.

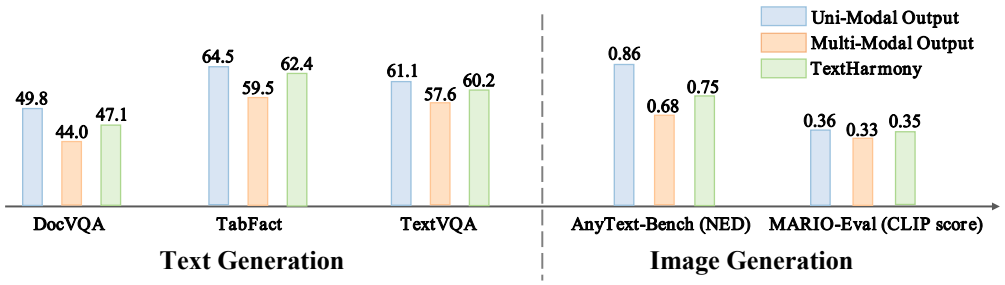


Figure 2: Comparison of single-modal and multi-modal output performance in text generation and image generation Tasks. "Uni-Modal Output" represents the results achieved by modality-specific supervised fine-tuning. "Multi-Modal Output" represents the results achieved by modal-independent supervised fine-tuning. Compared to the multi-modal output, a major performance degradation in the uni-modal output is observed for both text generation and image generation tasks.

Despite the strides made with Slide-LoRA in harmonizing comprehension and generation, the image quality produced by TextHarmony requires improvement. High-quality, detailed image caption data is crucial for visual text-centric image generation tasks. Thus, a detailed image caption dataset, DetailedTextCaps-100K, is created using an advanced closed-source MLLM with prompt engineering. By incorporating DetailedTextCaps-100K in the supervised fine-tuning stage, TextHarmony's image generation quality significantly improves over using original simple captions alone.

Utilizing the above approaches, TextHarmony is versatile in various visual text-centric tasks. In visual text perception tasks, TextHarmony is able to perform well in text detection and recognition, achieving state-of-the-art performance in text grounding tasks. In visual text comprehension tasks, TextHarmony achieves comparable performance to dedicated text comprehension models. In image generation tasks, TextHarmony matches the performance to dedicated visual text generation models. The contribution of this paper can be summarised in three folds:

- We introduce TextHarmony, a versatile large multimodal that allows for the unification of diverse visual text perception, comprehension, and generation tasks. TextHarmony performs comparably to specialized models in visual text perception, comprehension, generation, and editing.
- To mitigate the inconsistency between vision and language modalities in the generative space, we propose Slide-LoRA. Slide-LoRA dynamically aggregates modality-specific and modality-agnostic LoRA experts, partially decoupling the multimodal generative space. With Slide-LoRA, TextHarmony achieves comparable performance to modality-specific fine-tuning results by only adding 2% parameters in a singular model instance.
- A high-quality dataset of detailed visual text image captions (DetailedTextCaps-100K) is constructed with a closed-source MLLM to enhance the performance of visual text generation.

2 Related Work

2.1 Visual Text Comprehension

Recent multi-modal large language models have increasingly focused on comprehending images with textual information [26, 39, 15, 14, 72, 69, 41, 76, 60, 53, 67]. Among them, UniDoc [15] and mPLUG-DocOwl [72] creates noval text-oriented instruction-following datasets. mPLUG-DocOwl 1.5 [21] builds a parsing system for visual text and employs large-scale pre-training. Monkey [26], DocPedia [14], and HRVDA [27] comprehend dense text by supporting higher image resolution. TextMonkey [39] adopts shifted window attention and filters out significant tokens.

https://doi.org/10.52202/079017-3093

2.2 Visual Text Generation

Diffusion-based text-to-image generation has recently achieved impressive progress [19, 50, 77, 52, 48, 7], while the capability of rendering accurate, coherent text in images remains an open problem. DiffUTE [4] designs a text editing model through fine-grained glyph and position information control. GlyphControl [71] generates visual text images by first rendering the glyph and then performing the denoising process. TextDiffuer [6] builds a large-scale text images dataset with OCR annotations and generates visual text conditioned on character-level layouts. Further, TextDiffuer-2 [5] leverages a language model as the layout planner, relieving the character-level guidance. Anytext [66] integrates the text-control diffusion pipeline with an auxiliary latent module and a text embedding module, which has achieved remarkable success in multilingual visual text generation.

2.3 Unified Multi-Modal Comprehension and Generation

Current multi-modal large language models (M-LLMs) [82, 2, 1, 11] largely use predicting the next text token as the training objective but exert no supervision for visual data [57]. Recent researches [57, 56, 65, 80, 16, 17, 70] have been attempting to empower M-LLMs to generate visual elements. The Emu family [57, 56] learns with a predict-the-next-element objective in multi-modality and decodes the regressed visual embeddings by a visual decoder. MiniGPT-5 [80] introduces a fixed number of special tokens into the LLM's vocabulary as the generative tokens for images. SEED-LLaMA [16] proposes the SEED tokenizer, which produces discrete visual codes with causal dependency and high-level semantics. MM-Interleaved [65] extracts fine-grained visual details from multiple images' multi-scale feature maps, proving effective in generating interleaved image-text sequences. They all focus on generic multimodal generation, whereas no such work exists yet in the visual text domain.

3 Methodology

3.1 Model Architecture

Figure 3 presents an overview of TextHarmony. The backbone network follows the paradigm established by MM-Interleaved [65], where a Vision Encoder, an LLM, and an Image Decoder are internally integrated to empower the model to generate both visual and textual content. Specifically, the image embedding extracted by the Vision Encoder is abstracted by a Q-Former [25] and aligns with text tokens. Image and text tokens are then concatenated and forwarded through the LLM. The output token of the LLM either predicts text content or serves as the conditional input of image detokenization. The image decoder perceives the conditional input from LLM and generates images based on the denoising diffusion process [19].

Given the multi-modal input $\boldsymbol{X}$, the multi-modal generation task involves generating interleaved token sequences $\boldsymbol{Y}$, which can be detokenized into both image and text contents. The **token generation stage** is achieved by maximizing the conditional probability under the classic auto-regressive paradigm as follows:

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{l=1}^{L} p(\boldsymbol{Y}_l|\boldsymbol{X}, \boldsymbol{Y}_{<l}) = \prod_{l\in\mathcal{N}_T} p(\boldsymbol{Y}_l|\boldsymbol{X}, \boldsymbol{Y}_{<l}) \cdot \prod_{l\in\mathcal{N}_I} p(\boldsymbol{Y}_l|\boldsymbol{X}, \boldsymbol{Y}_{<l}), \quad (1)$$

where $\mathcal{N}_T$ and $\mathcal{N}_I$ refer to the index sets of the text and image tokens, respectively. $\boldsymbol{Y}_l$ is the l-th predicted token and $\boldsymbol{Y}_{<l}$ is the set of preceding tokens. After that, the text tokens are classified with a linear layer $q(\cdot)$ to produce text content, while the image tokens serve as the condition input in the denoising diffusion process to produce image content. The loss function in the **detokenization stage** consists of the above two parts:

$$\mathcal{L}(\boldsymbol{X}, \hat{\boldsymbol{Y}}) = \underbrace{-(\hat{\boldsymbol{Y}}_{l\in\mathcal{N}_T})^{\mathbf{T}} \cdot log[q(\boldsymbol{Y}_{l\in\mathcal{N}_T})]}_{\text{text generation}} + \underbrace{\mathbb{E}_{\epsilon,t}\|\epsilon - DM(\boldsymbol{Y}_{l\in\mathcal{N}_I}, t)\|^2}_{\text{image generation}}, \quad (2)$$

where $\hat{\boldsymbol{Y}}$ is the ground-truth token sequence and DM is the diffusion model for image denoising.

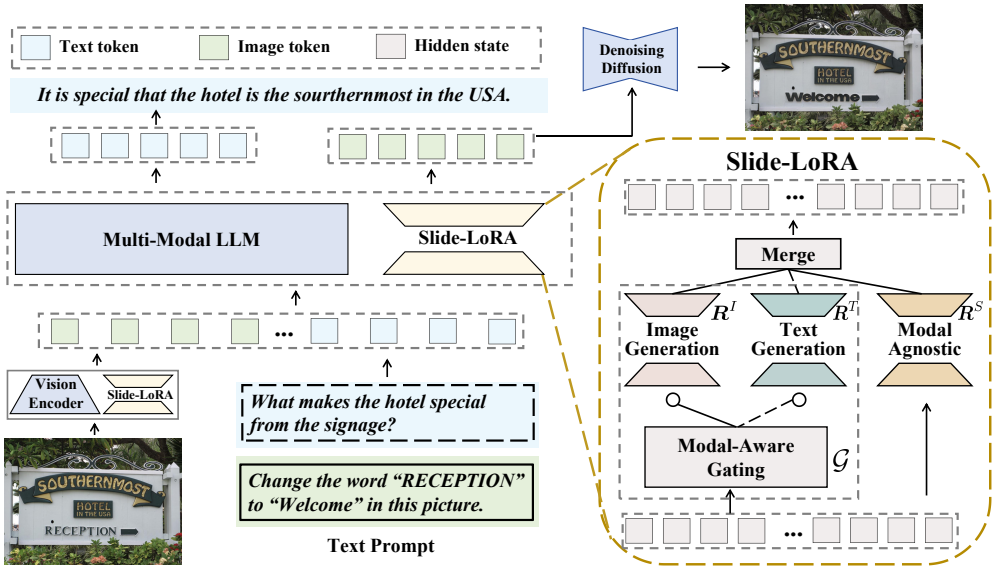


Figure 3: Pipeline of TextHarmony. TextHarmony generates both textual and visual content by concatenating a vision encoder, an LLM, and an image decoder. The proposed Slide-LoRA module mitigates the problem of inconsistency in multi-modal generation by partially separating the parameter space.

On the supervision of Equation 2, the optimization of TextHarmony is tremendously difficult due to inconsistent training objectives. Firstly, text generation aims to generate text from images, while image generation aims to generate images from text, which are mutually exclusive. Secondly, text generation is a classifier problem with a cross-entropy loss function, while image generation (*i.e.*, the denoising process) is a regression problem with a mean square error loss function. To this end, we propose to mitigate the training inconsistency problem by adaptively adjusting the forward pass according to the input tokens.

3.2 Slide-LoRA: Enhancing Image and Text Generation Consistency

To harmonize multi-modal generation tasks in a single model, the parameter space would be optimized for inconsistent or conflict training objectives, as stated before. We propose Slide-LoRA (which is shaped like a "slide" between different LoRA experts as shown in Figure 3), a novel module that can be conveniently inserted into Transformer layers as Low-Rank Adaptation (LoRA) [22] and introduces limited parameter increase. As such, Slide-LoRA spontaneously processes text generation and image generation in separate parameter spaces, thus relieving the inconsistent training problem.

As shown in Figure 3, Slide-LoRA is composed of a gating network $\mathcal{G}$ and three LoRA modules $(\boldsymbol{R}^T, \boldsymbol{R}^I, \boldsymbol{R}^S)$. $\boldsymbol{R}^T$ and $\boldsymbol{R}^I$ serve as the separate parameter space for text and image generation, respectively, while $\boldsymbol{R}^S$ aims to learn the knowledge shared by both text and image generation. Specifically, Given the input token sequence $\boldsymbol{x} \in \mathbb{R}^{L \times D}$, the gating network $\mathcal{G}$ (which can be established using either MLPs or rule-based discriminators) determines whether the processing of the input token sequence requires knowledge of text generation or image generation, and produces a scalar $\gamma = \mathcal{G}(\boldsymbol{x}) \in [0, 1]$. The output of the Slide-LoRA layer can be formulated as

$$\boldsymbol{O} = \frac{1}{2} \cdot \{[\gamma \geq 0.5] \cdot \boldsymbol{R}^T(\boldsymbol{x}) + [\gamma < 0.5] \cdot \boldsymbol{R}^I(\boldsymbol{x}) + \boldsymbol{R}^S(\boldsymbol{x})\}, \tag{3}$$

where $[\cdot]$ equals 1 if the condition inside is true and 0 otherwise. Slide-LoRA incorporates task-specific and task-shared knowledge from input tokens, thus separating the inconsistent training objective and learning the shared knowledge of text and image generation.

https://doi.org/10.52202/079017-3093

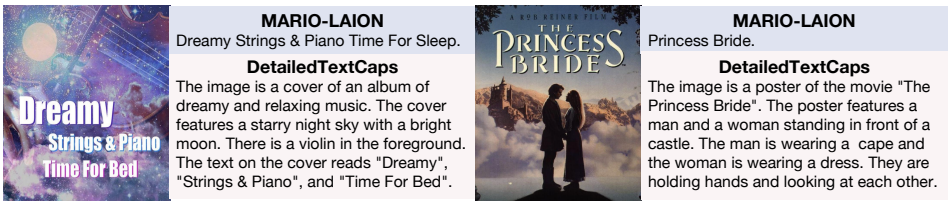


Figure 4: Captions from DetailedTextCaps-100K and MARIO-LAION for the same image. DetailedTextCaps-100K can better depict the textual elements in the image.

3.3 Multi-Modal Pre-Training and Comprehensive Fine-Tuning

TextHarmony training process consists of two stages. In the multi-modal pre-training stage, TextHarmony is trained on text-rich image-text corpus and learns to produce multi-modal outputs. In the comprehensive fine-tuning stage, we concurrently cultivate the text and image generation capabilities of TextHarmony by training on a series of text-centric tasks.

3.3.1 Stage 1: Multi-Modal Pre-Training

TextHarmony is pre-trained based on the pre-training weight of MM-Interleaved [65], with extra text-rich datasets including MARIO-LAION [6] and DocStruct4M [21]. MARIO-LAION contains 9M web images with brief captions and the according OCR results. DocStruct4M consists of 2M documents and 1M natural images with text-oriented structure annotations. We use MARIO-LAION for both text and image generation (*, i.e.*, either predict the caption of the image or generate the image based on the caption), and we use DocStruct4M for text generation only. In this stage, we freeze the vision encoder and the LLM, training only the Q-Former and the image decoder to obtain basic image understanding and generation capabilities.

3.3.2 Stage 2: Comprehensive Fine-Tuning

We integrate various text-centric datasets and employ uniform instructions for all tasks. In this stage, the vision encoder, Q-former, image decoder, and the proposed Slide-LoRA are trained to enhance the multi-modal generation and human-instruction-following capabilities of TextHarmony.

Visual Text Generation. In this task, TextHarmony generates images according to the text description and is required to render accurate and coherent text. Although MARIO-LAION contains captions of text-rich images, the description is oversimplified and lacks concentration on the textual elements within the image. To this end, we sample 100K images from MARIO-LAION and generate detailed captions about them, termed DetailedTextCaps-100K. The captions focus on both visual and textual elements in the images. This is achieved by prompting Gemini Pro [12], a pioneer multi-modal large language model, to generate detailed descriptions based on the sampled image and the OCR results. As shown in Figure 4, the image description from DetailedTextCaps-100K is more comprehensive compared with MARIO-LAION and can better depict the textual elements in the image.

Visual Text Editing. In this task, TextHarmony substitutes or renders text in the given location of the image and keeps the background consistent. We randomly mask the image with the help of MARIO-LAION's OCR results and fine-tune TextHarmony in a self-supervised manner.

Visual Text Comprehension. We employ the training set collected by Monkey [26] for the text-centric VQA fine-tuning. The training set involves 1.4M QA pairs and covers various text-rich scenarios.

Visual Text Perception. For the basic OCR capabilities, we randomly sample 1M images from MARIO-LAION and leverage the OCR annotations.

4 Experiment

4.1 Experimental Setup

Implementation details. We use CLIP-ViT-L/14 [49], Vicuna-13B [81], and Stable Diffusion v2.1 [51] as the vision encoder, the LLM and the image decoder, following MM-Interleaved [65]. The image resolution is increased to 896 to capture fine-grained features better. A Q-Former with 12 blocks is adopted to reduce the number of visual tokens to 512. In the multi-modal pre-training stage, the initial learning rate is set to $1e-5$, while in the fine-tuning stage, it is reduced to $5e-6$. The pre-training stage takes 3264 A-100 hours with a batch size of 256; While the fine-tuning stage takes 2352 A-100 hours with a batch size of 64.

Datasets and Metrics. We evaluate TextHarmony on a broad range of vision-language tasks. Visual Text Comprehension includes Document-Oriented VQA (InfoVQA [43], DocVQA [44], ChartQA [42]), Table VQA (TabFact [8], WTQ [47]), Scene Text-Centric VQA (TextVQA [55], OCRVQA [45], STVQA [3]) and OCRBench [38]. We adopt the Accuracy metric following TextMonkey [39]. Visual Text Generation includes AnyText-benchmark-EN [66] and MARIOEval [6], where the metric of NED, FID, and CLIP Score are used following AnyText [66] and TextDiffuser [6].

4.2 Quantitative Analysis

Table 1: Results of visual text comprehension. TextHarmony is compared with both uni-modal generation models and multi-modal generation models. We employ the Accuracy metric for all methods. TextHarmony* is trained without Slide-LoRA.

Method	Document-Oriented VQA			Table VQA		Scene Text-Centric VQA			OCRBench
	InfoVQA	DocVQA	ChartQA	TabFact	WTQ	TextVQA	OCRVQA	STVQA	
Models for text generation only									
LLaVAR [78]	16.5	12.3	12.2	-	-	41.8	24	39.2	346
UniDoc [15]	14.7	7.7	10.9	-	-	46.2	36.8	35.2	-
DocPedia [14]	15.2	47.1	46.9	-	-	60.2	57.2	45.5	-
mPLUG-Owl2 [74]	18.9	17.9	19.4	-	-	53.9	58.7	49.8	366
LLaVA1.5-7B [37]	14.7	8.5	9.3	-	-	38.7	58.1	38.1	297
Monkey [26]	25.8	**50.1**	**54.0**	49.8	25.3	**64.3**	**64.4**	54.7	514
InternVL [10]	23.6	28.7	45.6	-	-	59.8	30.5	**62.2**	**517**
InternLM-XComposer2 [13]	28.6	39.7	51.6	62.3	**28.7**	62.2	49.6	59.6	511
Models for both text and image generation									
SEED-LLaMA-14B [16]	23.5	6.5	13.8	49.2	13.2	14.4	16.3	20.1	357
MiniGPT5 [80]	2.1	1.6	1.4	4.9	0.9	2.8	2.3	2.4	68
MM-Interleaved [65]	17.0	8.1	11.9	40	15.1	37.2	11.7	26.4	197
TextHarmony*	26.1	44	36.2	59.5	26.1	57.6	51.9	47.2	397
TextHarmony-Chat	**28.9**	49.8	38.8	**64.5**	28.3	61.1	57.6	51.3	448
TextHarmony	28.5	47.1	38.0	62.4	27.1	60.2	55.3	49.7	440

Table 2: Text grounding performance on MARIO-Eval. The Acc@0.5 metric is employed.

TGDoc [68]	DocOwl 1.5 [21]	TextHarmony
82.5	84.3	**88.7**

4.2.1 Visual Text Comprehension and Perception

Comparison to Text-Generation and Multi-Modal Generation Methods. As shown in Table 1, we evaluate TextHarmony on a broad range of visual text comprehension tasks following [39]. As we can see, the performance of TextHarmony is comparable to that of SOTA methods specialized for visual comprehension overall. Specifically, on InfoVQA and TabFact, TextHarmony achieves one of the top performances, with an accuracy of 28.5% and 62.4%. On DocVQA, WTQ and TextVQA, TextHarmony scores 47.1%, 27.1% and 60.2%, which is competitive against top-tier methods like Monkey (50.1%, 25.3% and 64.3%) and InternLM-Xcomposer2 (39.7%, 28.7% and 62.2%). On ChartQA, STVQA, and OCRBench, our model lags behind Monkey, InternVL, and InternLM-XComposer2, which can be attributed to pre-training weights or training data variations. However, TextHarmony surpasses other methods like mPLUG-Owl2 and LLaVA1.5-7B. Furthermore, the performance of state-of-the-art multimodal generative models has a significant performance gap compared to TextHarmony.

https://doi.org/10.52202/079017-3093

Comparison to TextHarmony* and TextHarmony-Chat. To further validate the effectiveness of the proposed method, we train two copies of TextHarmony. TextHarmony* is trained without the proposed Slide-LoRA module, while TextHarmony-Chat is trained with only Visual Comprehension data and forms the upper bound of TextHarmony on visual text comprehension tasks. As shown in the bottom of Table 1, the performance of TextHarmony* is strictly lower than that of TextHarmony. For example, TextHarmony scores 3.1% higher on DocVQA, 2.9% higher on TabFact, and 3.4% higher on OCRVQA. Overall, the performance increase brought by Slide-LoRA on comprehension tasks is 2.5%. And the performance gap between TextHarmony and TextHarmony-Chat is thus lowered from 3.96% to 1.5%.

Comparison on Text Grounding and Recognition. As shown in Table 2, we evaluate the text grounding capabilities by sampling 300 visual text-position pairs from MARIO-Eval. TextHarmony achieves 88.7% and surpasses current MLLMs with text grounding capabilities (*i.e.,* TGDoc and DocOwl 1.5). The performance on OCRBench (Table 1) reflects the text recognition capabilities of MLLMs. Specifically, TextHarmony surpasses LLaVAR, mPLUG-Owl2, and LLaVA1.5-7B by 94, 74, and 143, while the gap between TextHarmony and top-tier MLLMs remains.

Table 3: Results of visual text editing and generation. TextHarmony is compared with both uni-modal generation models and multi-modal generation models. TextHarmony* is trained without Slide-LoRA.

	NED (↑)	FID (↓)	CLIP Score (↑)
Models for image generation only			
TextDiffuser-2 [6]	0.81	336	0.35
GlyphControl [71]	-	345	**0.36**
AnyText [66]	**0.88**	352	**0.36**
Models for both text and image generation			
SEED-LLaMA-14B [16]	0.11	348	0.27
MiniGPT5 [80]	0.02	380	0.25
MM-Interleaved [65]	0.04	412	0.29
TextHarmony*	0.68	356	0.33
TextHarmony-Gen	0.86	**330**	**0.36**
TextHarmony	0.75	342	0.35

4.2.2 Visual Text Generation and Editing

As shown in Table 3, we compare TextHarmony with diffusion model-based text-to-image generation methods and multi-modal generation models. We evaluate the performance of visual text editing by sampling 200 images from AnyText-benchmark-EN and randomly choosing one available text polygon in each image. We use the NED metric to evaluate visual text editing. The performance of Image Generation is evaluated by sampling 100 text-image pairs from MARIO-Eval using the FID and CLIP score metrics. TextHarmony achieves 0.75 on NED and 0.35 on CLIP score, which is comparable to GlyphControl and TextDiffuser. Note that TextHarmony is a multi-modal generation model that is not specialized for image generation and editing, while other multi-modal generation models can hardly generate visual text.

We further train an image generation version of HarmonyText with only image generation/editing data, termed HarmonyText-Gen. As we can see, HarmonyText-Gen achieves 0.86 on NED and 0.36 on CLIP score, surpassing most image-generation methods and achieving comparable results to Anytext. Moreover, the application of Slide-LoRA brings an improvement of 0.07 on NED (from 0.68 to 0.75) and 0.02 on CLIP score (from 0.33 to 0.35), further validating the effectiveness of Slide-LoRA.

4.3 Ablation Studies

Impact of the Config Choice of Slide-LoRA. As shown in Table 4, we compare the performance of TextHarmony by varying the total number of LoRA modules used in Slide-LoRA. As we can see, the varied number of LoRA modules has little influence on comprehension and generation performance. For example, the accuracy on TextVQA is 60.2%, 60.4%, and 60.4% when n is increased from 3 to 9.

Table 4: Ablation studies of the config choices of Slide-LoRA and the places to insert Slide-LoRA.

		Image to Text			Text to Image	
		TextVQA	InfoVQA	OCRBench	NED	CLIP Score
Config Choice of Slide-LoRA	w/o Slide-LoRA	57.6	26.1	426	0.68	0.33
	n=3, s=1	60.2	28.5	440	0.75	0.35
	n=6, s=2	60.4	28.2	440	0.73	0.35
	n=9, s=3	60.4	28.3	442	0.74	0.36
Place to Insert Slide-LoRA	Vision Encoder	58	26.7	432	0.69	0.34
	LLM	59.9	28.1	434	0.73	0.35
	Both	60.2	28.5	440	0.75	0.35

Table 5: Ablation studies of DetailedTextCaps.

DetailedTextCaps-100K	NED	FID (↓)	CLIP Score
w/o	0.70	368	0.32
w/	0.75	342	0.35

Impact of varied places to insert Slide-LoRA. As shown in Table 4, simply inserting Slide-LoRA into the vision encoder would strongly decrease the performance. On the contrary, simply inserting Slide-LoRA into the LLM brings limited performance degradation. The LLM processes multi-modal inputs and generates both visual and textual tokens, while the vision encoder receives only image inputs and generates visual tokens. As a result, when inserted only in the vision encoder, Slide-LoRA would be unable to separate parameter spaces for text generation and image generation.

Impact of DetailedTextCaps-100K. We validate the effectiveness of the proposed DetailedTextCaps-100K dataset in Table 5. By removing DetailedTextCaps-100K from the training set and using the original caption from MARIO-LAION, the performance of both the image editing and the image generation decreased. Specifically, the NED decreased from 0.75 to 0.70, and the CLIP score decreased from 0.35 to 0.32. This further validates the effectiveness of detailing the image caption in visual text when training TextHarmony.

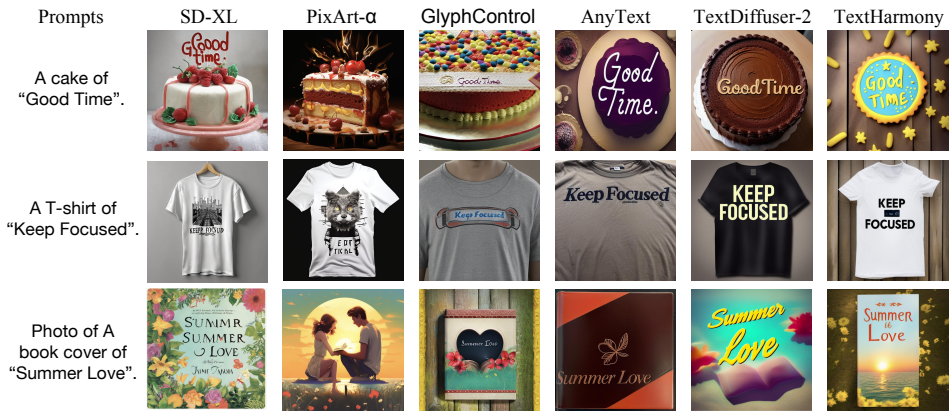


Figure 5: Visualisation of visual text generation.

4.4 Qualitative Analysis

We present examples of visual text generation in Figure 5 and examples of visual text editing of TextHarmony in Figure 6.

5 Limitation

Despite the strides made by TextHarmony in unifying visual text generation and comprehension within a single model instance, it is important to acknowledge certain limitations of our approach.

https://doi.org/10.52202/079017-3093

Figure 6: Visualisation of visual text editing.

Firstly, the performance of TextHarmony in visual text perception and comprehension tasks marginally lags behind some of the state-of-the-art open-source models. This discrepancy could be attributed to variations in the training data or the foundational model itself. Secondly, the model's proficiency in generating images is less robust in scenarios characterized by dense text, indicating an area that necessitates further investigation and enhancement. These limitations highlight the need for ongoing research to refine the capabilities of TextHarmony and similar multimodal generative models.

6 Conclusion

In summary, this work presents TextHarmony, a versatile multimodal generative model adept at reconciling the disparate tasks of visual text comprehension and generation. Utilizing the proposed Slide-LoRA mechanism, TextHarmony synchronizes the generation process of both vision and language modalities within a singular model instance, effectively addressing the inherent inconsistencies between different modalities. The model architecture is proficient in executing tasks that involve processing and generating images, masks, texts, and layouts, particularly within the realms of Optical Character Recognition (OCR) and document analysis. The accomplishments of TextHarmony herald the significant potential for comprehensive multimodal generative models within the visual text domain. The adaptability of TextHarmony indicates that models of a similar nature could be effectively employed across a diverse array of applications, offering the prospect of revolutionizing sectors that depend on the intricate interplay of visual text comprehension and generation.

Acknowledgement. This work is partially supported by National Natural Science Foundation of China (No.62222602, No.62106075, No.62176092, No.62302167, No.62476090, U23A20343), Natural Science Foundation of Shanghai (23ZR1420400), Shanghai Sailing Program (23YF1410500), Natural Science Foundation of Chongqing, China (CSTB2023NSCQ-JQX0007, CSTB2023NSCQ-MSX0137), Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (23CGA34).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019.

[4] Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Changhua Meng, Huijia Zhu, Weiqiang Wang, et al. Diffute: Universal text editing diffusion model. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] 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*, 2023.

[6] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[8] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

[9] Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. Octavius: Mitigating task interference in mllms via moe. *arXiv preprint arXiv:2311.02684*, 2023.

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] DeepMind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. 2023.

[13] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[14] Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023.

[15] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*, 2023.

[16] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[17] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[18] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023.

 https://doi.org/10.52202/079017-3093

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

[20] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023.

[21] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[23] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.

[24] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20543–20554, 2023.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[26] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023.

[27] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. Hrvda: High-resolution visual document assistant. *arXiv preprint arXiv:2404.06918*, 2024.

[28] Daizong Liu, Xiang Fang, Pan Zhou, Xing Di, Weining Lu, and Yu Cheng. Hypotheses tree building for one-shot temporal sentence localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1640–1648, 2023.

[29] Daizong Liu, Yang Liu, Wencan Huang, and Wei Hu. A survey on text-guided 3d visual grounding: Elements, recent advances, and future directions. *arXiv preprint arXiv:2406.05785*, 2024.

[30] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. Saanet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020.

[31] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1665–1673, 2022.

[32] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. *arXiv preprint arXiv:2109.06398*, 2021.

[33] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[34] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020.

[35] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1683–1691, 2022.

[36] Daizong Liu, Pan Zhou, Zichuan Xu, Haozhao Wang, and Ruixuan Li. Few-shot temporal sentence grounding via memory-guided semantic learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5):2491–2505, 2022.

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

[38] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[39] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

[40] Yuliang Liu, Jiaxin Zhang, Dezhi Peng, Mingxin Huang, Xinyu Wang, Jingqun Tang, Can Huang, Dahua Lin, Chunhua Shen, Xiang Bai, and Lianwen Jin. Spts v2: Single-point scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15665–15679, 2023.

[41] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. *arXiv preprint arXiv:2404.05225*, 2024.

[42] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[43] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[44] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[45] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[46] OpenAI. Gpt-4v(ision) system card. 2023.

[47] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.

[48] 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. *arXiv preprint arXiv:2307.01952*, 2023.

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

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[53] Bin Shan, Xiang Fei, Wei Shi, An-Lan Wang, Guozhi Tang, Lei Liao, Jingqun Tang, Xiang Bai, and Can Huang. Mctbench: Multimodal cognition towards text-rich visual scenes benchmark. *arXiv preprint arXiv:2410.11538*, 2024.

[54] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[55] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

 https://doi.org/10.52202/079017-3093

[56] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.

[57] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023.

[58] Jingqun Tang, Weidong Du, Bin Wang, Wenyang Zhou, Shuqi Mei, Tao Xue, Xing Xu, and Hai Zhang. Character recognition competition for street view shop signs. *National Science Review*, 10(6):nwad141, 2023.

[59] Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, et al. Textsquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*, 2024.

[60] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.

[61] Jingqun Tang, Wenming Qian, Luchuan Song, Xiena Dong, Lan Li, and Xiang Bai. Optimal boxes: boosting end-to-end scene text recognition by adjusting annotated bounding boxes via reinforcement learning. In *European Conference on Computer Vision*, pages 233–248. Springer, 2022.

[62] Jingqun Tang, Su Qiao, Benlei Cui, Yuhang Ma, Sheng Zhang, and Dimitrios Kanoulas. You can even annotate text with voice: Transcription-only-supervised text spotting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4154–4163, 2022.

[63] Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. Few could be better than all: Feature sampling and grouping for scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4563–4572, June 2022.

[64] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023.

[65] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.

[66] 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.

[67] An-Lan Wang, Bin Shan, Wei Shi, Kun-Yu Lin, Xiang Fei, Guozhi Tang, Lei Liao, Jingqun Tang, Can Huang, and Wei-Shi Zheng. Pargo: Bridging vision-language with partial and global views. *arXiv preprint arXiv:2408.12928*, 2024.

[68] Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*, 2023.

[69] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.

[70] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[71] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[72] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

[73] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

[74] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

[75] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, and Xiang Bai. Turning a clip model into a scene text detector. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[76] Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Structextv2: Masked visual-textual prediction for document image pre-training. *arXiv preprint arXiv:2303.00289*, 2023.

[77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[78] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[79] Zhen Zhao, Jingqun Tang, Chunhui Lin, Binghong Wu, Can Huang, Hao Liu, Xin Tan, Zhizhong Zhang, and Yuan Xie. Multi-modal in-context learning makes an ego-evolving scene text recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15567–15576, June 2024.

[80] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.

[81] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[82] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

 https://doi.org/10.52202/079017-3093

A Appendix / supplemental material

A.1 Experiments and Visualisation

A.1.1 Construction of DetailedTextCaps-100K

DetailedTextCaps-100K is constructed by re-generating the captions about the images from MARIO-LAION through prompting Gemini Pro. Given a sampled image, Gemini Pro is required to generate a detailed caption about the image through the following prompt:

> *Play an image content analysis expert. First, analyze all the image contents in a comprehensive manner and pay special attention to the textual elements in this image. Then,* ***describe the image in as much detail as you can.***

After that, we eliminate the sample if its generated caption is longer than 100 or if the caption contains non-ASCII characters. Further, Advanced MLLMs, including Gemini, are utilized to check the quality of the generated captions through the following prompt:

> *Play an image content analysis expert. First, analyze all the image contents comprehensively and pay special attention to the textual elements in this image. Then,* ***judge whether the caption is consistent with the image and whether the caption can thoroughly describe the image.***

After the above filtering system, we obtained $102,421$ images with high-quality text-oriented captions.

To validate the effectiveness of DetailedTextCaps-100K before merging it into the training set, we let GPT-4V [46] determine which is better, the captions from DetailedTextCaps-100K or the captions from MARIO-LAION. Specifically, we randomly sample 88 image-caption pairs from DetailedTextCaps, and GPT-4V is asked to answer the following question for each image:

> *Which is a better description for this picture? A:<Caption from DetailedTextCaps>, B:<Caption from MARIO-LAION>. Please only answer A or B without any other choices.*

The evaluation result is $82:6$, indicating that GPT-4V chooses captions from our DetailedTextCaps-100K as the better image description in most cases. The above evaluation protocol may contain biases from GPT-4V. Thus, we present this result for reference. Nevertheless, the ablation studies conducted in Section 4.3 prove the effectiveness of DetailedTextCaps-100K in favoring visual text generation. More examples of DetailedTextCaps-100K are shown in Figure 7.

Figure 7: More Examples of DetailedTextCaps-100K.

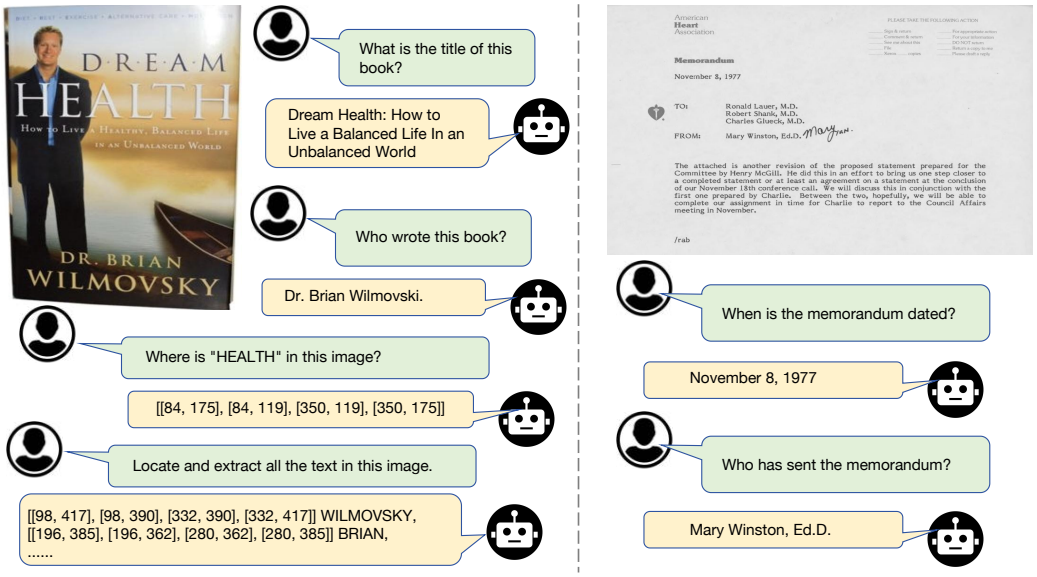


Figure 8: Visualisation of TextHarmony's visual text comprehension and perception capabilities.

A.1.2 Visualisation of visual text comprehension and perception

Figure 8 presents examples showing the visual text comprehension and perception capabilities of TextHarmony.

Table 6: Prompt design in comprehensive fine-tuning.

Task	Prompt
Visual Text Perception	What is the text in <mask> in this image? Where is <text> in this image? Extract all the text in this image. Locate all the text in this image. Locate and extract all the text in this image.
Visual Text Generation	Generate an image according to the caption.
Visual Text Editing	Fill the masked part in this image with <text>

A.1.3 Prompt design and data formatting in comprehensive fine-tuning

Table 6 presents the prompt design of TextHarmony in the stage of comprehensive fine-tuning. For visual text comprehension and perception, the input data sequence in the training stage is formulated as:

Answer the following question based on the image. <Image> Question: <Question> Answer: <Answer>.

For visual text generation and editing, the input data sequence in the training stage is formulated as follows:

<Image> <Instruction> <Target Image>.

We employ an all-black image as the input <Image> for visual text generation to make the data formatting consistent with visual text editing. The input data is processed into token sequences by image and text tokenizer, and TextHarmony is trained in the classical auto-regressive manner.

https://doi.org/10.52202/079017-3093

Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. We summarize these contributions in the conclusion of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper conducts a critical evaluation of its methodology and acknowledges its inherent limitations. We summarize the limitations at the end of the main text.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: -

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides all necessary details to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-3093

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the necessary details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: -

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources required to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conduct in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-3093

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of assets used in the paper are properly credited. The license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are well documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/079017-3093

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.