UNIT: Unifying Image and Text Recognition in One Vision Encoder

Yi Zhu[1], **Yanpeng Zhou**[1], **Chunwei Wang**[1], **Yang Cao**[2], **Jianhua Han**[1], **Lu Hou**[1], **Hang Xu**[1]

[1]Huawei Noah's Ark Lab, [2]Hong Kong University of Science and Technology

`https://github.com/yeezhu/UNIT`.

Abstract

Currently, vision encoder models like Vision Transformers (ViTs) typically excel at image recognition tasks but cannot simultaneously support text recognition like human visual recognition. To address this limitation, we propose **UNIT**, a novel training framework aimed at **UN**ifying **I**mage and **T**ext recognition within a single model. Starting with a vision encoder pre-trained with image recognition tasks, UNIT introduces a lightweight language decoder for predicting text outputs and a lightweight vision decoder to prevent catastrophic forgetting of the original image encoding capabilities. The training process comprises two stages: *intra-scale pretraining* and *inter-scale finetuning*. During intra-scale pretraining, UNIT learns unified representations from multi-scale inputs, where images and documents are at their commonly used resolution, to enable fundamental recognition capability. In the inter-scale finetuning stage, the model introduces scale-exchanged data, featuring images and documents at resolutions different from the most commonly used ones, to enhance its scale robustness. Notably, UNIT retains the original vision encoder architecture, making it **cost-free** in terms of inference and deployment. Experiments across multiple benchmarks confirm that our method significantly outperforms existing methods on document-related tasks (e.g., OCR and DocQA) while maintaining the performances on natural images, demonstrating its ability to substantially enhance text recognition without compromising its core image recognition capabilities.

1 Introduction

Vision encoder models [6, 43, 44, 53, 41, 24] are crucial for extracting high-level visual features, thereby enhancing performance across a range of downstream tasks [9, 38, 8, 14, 57]. They play a vital role in integrating visual information into intelligent applications, driving advancements in both computer vision and artificial intelligence. In recent years, Vision Transformers (ViTs)based encoder models [16, 66, 27, 58, 71] have revolutionized the field of computer vision. ViT models pretrained on image classification tasks are extensively used as backbones for a wide array of image recognition tasks and achieves state-of-the-art performances. Another line of remarkable ViT models are trained via image-text cross-modal contrastive learning [43, 64, 52, 67]. These models are often used as plug-in vision encoders in Large-scale Vision-Language Models (LVLMs), which have emerged as versatile tools with the potential to revolutionize various domains, including healthcare diagnostics, autonomous driving, digital assistants, and advanced content analysis in media and entertainment. Despite these advancements, existing vision encoder models, exhibit a significant limitation: they have not yet demonstrated *the capability to simultaneously support both image and text recognition* like humans do. Such ability is essential for document analysis applications, for instance, a model must accurately recognize and interpret textual information embedded within complex layouts, such as tables, graphs, and mixed media.

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

Text recognition [50, 49, 12], particularly in the context of dense documents and complex visual environments, presents unique challenges that are distinct from those encountered in pure image recognition. While image recognition typically focuses on global feature extraction and classification, text recognition requires precise local feature extraction and sequence prediction. ViTs, though adept at handling global image features, often struggle with the detailed local features needed for accurate text recognition, particularly in high-resolution document images where fine details are crucial. A straightforward solution is to finetune pre-trained ViTs using high-resolution documents. However, this approach requires interpolating the pre-trained positional embeddings to handle longer sequences. Such interpolation with a large magnification will disrupt the alignment between the original positional embeddings and their corresponding spatial positions, thereby impacting the model's performance.

Existing methods [39, 4, 5] build document-specific models by retraining ViTs exclusively on the Optical Character Recognition (OCR) task, thereby discarding the original image encoding capability. In downstream applications, these models are often ensembled with other expert models, requiring users to specify the data type in advance, which is inadequate for scenarios requiring dynamic and simultaneous recognition of both image and text without prior knowledge of the input content. Other ensemble-based methods [53, 54] use similar model architectures trained independently on image and text recognition tasks. Inputs are fed into both models, and the output features are concatenated to form a unified representation. However, this approach will result in a significant increase in computational cost. Additionally, some LVLM methods [62, 63] enhance document analysis tasks (e.g., DocQA) in an OCR-free manner by finetuning with document-instruction data. Nevertheless, their capabilities are limited to handling images with prominent text and fail with dense documents, struggling to generalize across different font sizes, typefaces, and backgrounds.

In this paper, we propose **UNIT**, a novel training framework for **UN**ifying **I**mage and **T**ext recognition abilities within a single model. UNIT upgrades an existing Vision Transformer (ViT) model to effectively integrate text recognition. First, a lightweight language decoder (e.g., OPT-125M) is introduced to decode the learned visual features into text sequences in an auto-regressive manner, enabling the encoder model to capture fine-detailed shape and sequential information necessary for text recognition. Then, a tiny vision decoder (e.g., two MLP layers) is introduced to reconstruct the visual features of the original vision encoder from the newly learned features, preventing catastrophic forgetting of the model's fundamental encoding ability for natural images. The training pipeline involves two stages. The first is the **intra-scale pretraining** stage, where the model takes images and documents at their commonly used resolutions as inputs, specifically *low-resolution images and high-resolution documents* ($\times 4$ times the low-resolution). During this stage, the model is optimized with three objectives: OCR for document inputs, feature reconstruction of the original ViT encoder's features and image captioning for natural image inputs. The second stage is **inter-scale finetuning** stage, where the model is trained under a scale-exchanged setting with *high-resolution images and low-resolution documents*, enhancing its scale robustness for both image and text recognition. It is important to note that UNIT retains the original vision encoder architecture, ensuring that there is no increase in inference cost.

Extensive experiments show that UNIT significantly outperforms document-specific models on OCR tasks while maintaining core image recognition capabilities. When integrated into LVLMs, it also benefits downstream document analysis tasks without degrading image understanding. This demonstrates that UNIT effectively unifies image and text recognition abilities.

The main contributions of this paper are summarized as follows:

- We propose UNIT, a novel framework that unifies image and text recognition in a single vision encoder through joint training of multi-tasks. The model retains the original vision encoder architecture, ensuring cost-free deployment and no increase in inference cost.

- The training paradigm comprises two key stages: an intra-scale pretraining stage, where images and documents are trained at their commonly used resolutions, and an inter-scale finetuning stage, where their resolutions are exchanged to enhance scale robustness.

- Extensive experiments demonstrate that UNIT significantly enhances performance on vision tasks requiring text recognition compared to their counterparts, while preserving the fundamental encoding ability on natural images.

2 Related Work

Vision Transformer for Text Recognition. In recent work, the application of Vision Transformers (ViT) to document reading has gained traction due to their success in image recognition [30, 10, 24, 4, 15]. Nougat [5]leverages a ViT model for Optical Character Recognition (OCR) tasks, focusing on converting human-readable scientific documents into a machine-readable markup language. Similarly, KOSMOS-2.5 [19] employs a ViT-based vision encoder coupled with a Transformer-based language decoder, aiming to serve as a versatile tool for diverse text-intensive image understanding tasks through supervised fine-tuning. The Donut model [23], while introducing an OCR-free transformer for visual document understanding, may encounter challenges in generalizing to unfamiliar document types. Despite their demonstrated efficacy in specialized OCR scenarios, these models may not fully harness the original image encoding capabilities present in pre-trained ViT architectures. This limitation can restrict their applicability to text-only images.

LVLMs for Document Analysis. Several recent studies have explored OCR-free visual-situated language understanding using Large Vision-Language Models (LVLMs) [47, 55, 59, 1, 2, 70, 40, 28], where document-instruction data is used to fine-tune LLMs without adapting the vision encoder [62, 60, 61]. These methods often rely on a pretrained Vision Transformer (ViT) as a fixed vision encoder, which may present several challenges: 1) The frozen ViT may limit text recognition capabilities, especially when documents feature varied fonts, styles, or low image quality not seen during its pretraining. 2) The model's reliance on a constant image resolution can lead to inaccuracies in text recognition for high-resolution documents or those with noise and distortion. Additionally, several recent works [53, 54] have successfully enhanced LVLMs with OCR-like abilities, by retraining a document-specific Vision Transformer (ViT) and then concatenating its output visual features with those from a pretrained ViT model. However, this method comes with trade-offs, notably an increase in computational expense and a significant underutilization of the model's capacity, leading to inefficiencies. In contrast, our UNIT integrates both capabilities within a single model, offering greater efficiency, cost-free deployment, and no increase in inference time.

Multi-scale Vision Transformer. Vision Transformer models are commonly pre-trained to process images at a fixed resolution, such as 224 or 336 pixels, which is commonly used for many image understanding tasks [17, 20, 33, 26, 25]. However, these resolutions might be insufficient for precisely discerning tightly packed text in higher-resolution documents. Consequently, a multi-scale training approach for Vision Transformers is crucial for effectively managing both image and text recognition challenges. The progression of Vision Transformer architectures has embraced a notable trend toward multi-scale modeling. This paradigm enables the models to capture and process visual information across different scales, enhancing their capability to understand complex scenes and texts. Some methods [29, 51, 56] integrate a hierarchical pyramid structure into the original Vision Transformer architectures, which continue to operate at a constant resolution. In contrast to these methods, our proposed UNIT model breaks the limitations of fixed-resolution inputs. UNIT is designed to handle multi-scale training for multi-tasks, yielding a unified representation with scale robustness for both images and documents.

3 Methodology

3.1 Preliminaries

Backbone. We employ the Vision Transformer (ViT) [16] architecture as our backbone. Let $f_\theta(\cdot)$ be the vision encoder with parameters θ, for an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W represent the image height and width, respectively. ViT model first partitions each image into fixed-size of patches in $p \times p$ pixels, and then encode these patches into hidden features of dimension d. Subsequently, the features are added with their corresponding position embeddings and fed into multiple layers of stacked Transformer blocks, interacting with each other via attention mechanisms. Finally the model outputs visual tokens $\mathbf{X} \in \mathbb{R}^{N \times d}$, as:

$$\mathbf{X} = f_\theta(\mathbf{I}) = \{\mathbf{x}_i\}_{i=1}^{N}, \tag{1}$$

where $N = N_s + N_c$ denotes the total number of visual tokens, comprising the spatial tokens $N_s = \frac{H}{p} \times \frac{W}{p}$ and the number of [CLS] tokens N_c.

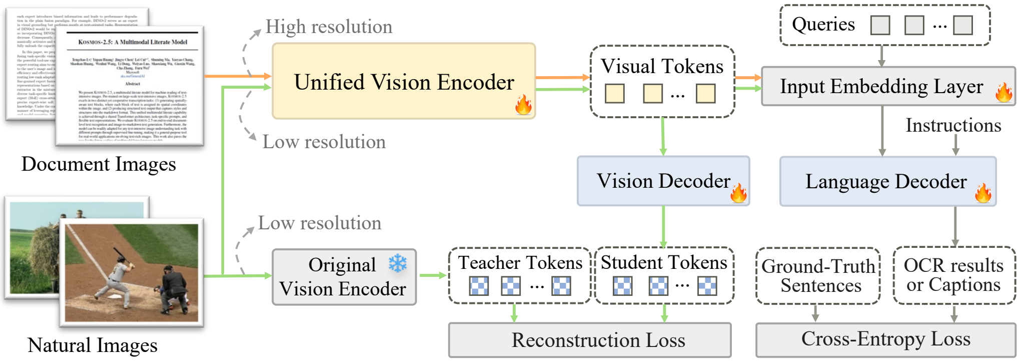


Figure 1: Overview of UNIT Architecture. The model processes high-resolution documents and low-resolution images, generating a set of visual tokens. These tokens pass through an input embedding layer, with document tokens fed into the language decoder to predict text sequences, enhancing the model's text recognition capability. Simultaneously, to preserve the model's original image encoding ability, the visual tokens from natural images are reconstructed via a lightweight vision decoder, mimicking the output of the teacher model. Additionally, an image captioning task is included alongside the OCR task to further enhance image understanding.

Architecture. UNIT aims to unify image and text recognition capabilities within a single ViT model, as illustrated in Fig. 1. This is achieved through the introduction of a lightweight language decoder (e.g., OPT-125M) for document-level OCR, enabling the ViT model to recognize text. Additionally, to preserve the original image recognition capabilities of the vision encoder, UNIT incorporates a vision decoder responsible for token-wise feature reconstruction from the original ViT model. To further enhance image understanding, an image captioning task is integrated alongside the OCR task. UNIT is trained in a multi-scale setting, where images and texts are processed at their commonly used resolutions during pretraining and then finetuned at other resolutions.

3.2 Text Recognition Ability Enhancement

Multi-Scale Inputs. In contrast to image recognition that focuses on global feature extraction and classification, text recognition requires fine local feature extraction and sequence prediction. ViTs are excellent at global feature extraction but may not perform well in text recognition. The reason is that ViTs use a learned position embedding for each input patch in an image, which in turn forces that the model always operate at a constant resolution, typically chosen as 224, 256 or 336. At these resolutions, the generated visual features may lose critical information, leading to difficulties in recognizing characters from dense texts such as PDF documents, websites, or digital books. Naively cropping high-resolution document images into low-resolution inputs would break the sequential order of characters and may impair the recognition of letters at the cutting boundaries. To prevent these issues, we introduce the Cropped Position Embedding (CPE) [22] augmentation, which interpolates the original position embeddings to match the number of positions of the maximum input size. For low-resolution images, the position embeddings are then randomly cropped and interpolated to match the number of input patches for the original model.

Language Decoder. Conventional methods often employ CLIP-style contrastive learning to align images with their language annotations. However, this approach presents two notable drawbacks in our context. Firstly, contrastive learning relies on a CLIP pretrained text encoder to generate language embeddings. The maximum sequence length of 77 tokens is insufficient for text recognition annotations, especially in dense documents. Secondly, such training paradigm is inadequate for the vision encoder to accurately capture each word, leading to undesirable information loss during the encoding process. Taking the above factors into account, we introduce a lightweight Transformer decoder, e.g., OPT-125M [69], to efficiently predict the language sequences presented in the input documents. Similar to LVLMs, we employ an input embedding layer to project the visual features from the vision encoder into the embedding space of the language decoder. Here we initialize the input embedding layer with a two-layer Q-Former [25] with K learnable query tokens $\mathbf{Q} = \{\mathbf{q}_k\}_{k=1}^{K}, \mathbf{q}_k \in \mathbb{R}^d$. The visual tokens $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ generated from the vision encoder are first fed into the input embedding layer and then the language decoder. We denote this process as $f_\phi(\cdot)$ and ϕ

denotes the parameters of the input embedding layer and the language decoder:

$$\{\mathbf{z}_1, ..., \mathbf{z}_N\} = f_\phi(\{\mathbf{x}_1, ..., \mathbf{x}_N\}, \{\mathbf{q}_1, ..., \mathbf{q}_K\}). \tag{2}$$

The decoder uses the last hidden state $\mathbf{z}_t^L \in \mathbb{R}^d$ at time step t to predict language sequence in the form of text tokens $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}$ via the language decoder in an auto-regressive manner. The cross-entropy loss for training the auto-regressive Transformer decoder is defined as follows:

$$\mathcal{L}_{\text{lan}}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{t=1}^{T} \log P(\hat{y}_t = y_t | \mathbf{y}_{1:t-1}, \mathbf{z}_t^L), \tag{3}$$

where $\mathbf{y} = \{y_1, y_2, ..., y_T\}$ represents the target sequence, and the probability $P(\hat{y}_t = y_t | \mathbf{y}_{1:t-1}, \mathbf{z}_t^L)$ is obtained from the softmax output of the final layer of the decoder.

3.3 Image Recognition Ability Preservation

Vision Decoder. Finetuning the vision encoder exclusively on document datasets would result in severe catastrophic forgetting of its original image encoding capabilities. To mitigate this, we introduce an token-wise feature reconstruction task on natural image datasets, ensuring that the newly learned vision encoder retains its ability to encode visual concepts. We incorporate a visual decoder $f_\pi(\cdot)$ consisting of two fully connected layers with a GeLU activation function in the intermediate layer. The decoder processes each visual token independently, preserving the positional information inherent in each token more effectively. We utilize the original pretrained ViT model as a teacher model to provide original features for each natural image, providing supervision signals for visual feature reconstruction. Denoting the teacher tokens as $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_i\}_{i=1}^N, \hat{\mathbf{x}}_i \in \mathbb{R}^d$, we enforce the alignment of the new learned student tokens with the original ones using a weighted sum of the cosine distance L_{cos} and smooth L1 loss L_{l1}. This approach ensures consistency in both vector direction and magnitude between the new and original features, as:

$$\mathcal{L}_{\text{vis}}(\mathbf{X}, \hat{\mathbf{X}}) = \sum_{i \in \mathcal{C}} L_{\cos}(f_\pi(\mathbf{x}_i), \hat{\mathbf{x}}_i) + \mu L_{\text{l1}}(f_\pi(\mathbf{x}_i), \hat{\mathbf{x}}_i), \tag{4}$$

where $\mathcal{C}$ represents a subset of features randomly sampled from the $\hat{\mathbf{X}}$, to mitigate overfitting and enhance robustness during training. μ denotes the loss weight scalar.

Language Decoder. Furthermore, we incorporate an image captioning task alongside the text recognition task to ensure that the learned visual tokens can be effectively projected into the language space, thereby enhancing image understanding ability. The forward process and loss formulation for image inputs are the same as those for document inputs.

3.4 Intra-Scale Pretraining

As described in Sec. 3.2 and Sec. 3.3, UNIT tackles images and documents at different input scales with different optimization objectives. In this section, we introduce the intra-scale pretraining stage (see Fig. 2a) where image and text recognition are trained at a fixed scale, namely, low resolution for images and high resolution for documents, without considering variations in scale.

Dataset Preparation. Let $\mathcal{D}$ be a curated dataset with 5M samples, where $\mathcal{D} = \mathcal{D}_{\times 1}^I \cup \mathcal{D}_{\times 4}^T$. Here, $\mathcal{D}_{\times 1}^I$ represents a dataset of natural images annotated with coarse captions (less than 30 words). The images are resized to $\times 1$ times the original scale r, primarily sourced from the Conceptual Caption dataset [46], comprising 3M samples. Meanwhile, $\mathcal{D}_{\times 4}^T$ represents a dataset of documents annotated with dense OCR data (more than 500 words). The document images are resized $\times 4$ times the original scale r, sourced from our synthetic dataset of English PDF documents, comprising 2M samples.

Instruction Prompts. The instruction prompts follow the format used in popular LVLMs [70, 25, 1, 3, 2], where image tokens are prefixed with text tokens. Specifically, we use two special tokens "<img>" and "</img>", to indicate the start and the end position of the image tokens, followed by instructions that indicate task requirements. For natural images, the prompt is set as "Give a caption of this image:" and the LLM outputs a language sequence summarizing the content of the image. For document images, the prompt is set as "Read the text in this image:" and the LLM outputs sentences row by row as they appear in the document in the order of reading.

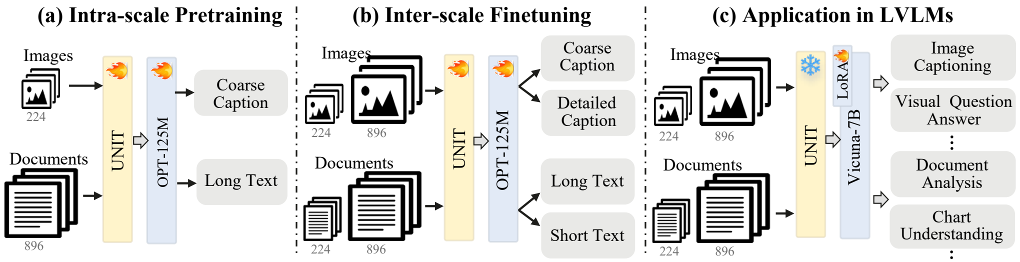


Figure 2: Illustration of the UNIT training paradigm. The (a) intra-scale pretraining stage processes images and documents at their commonly used resolutions to integrate basic text recognition with existing image recognition capabilities. The (b) inter-scale finetuning stage processes scale-exchanged data and tasks to enhance scale robustness, benefiting downstream document analysis tasks when integrated into (c) LVLMs applications.

Joint Optimization. We train all parameters, including the ViT backbone, vision decoder and language decoder, using images and documents annotated with language sentences. During training, the model is optimized by the cross-entropy loss $\mathcal{L}_{\text{lan}}$ between language outputs and annotations, and also the feature reconstruction loss $\mathcal{L}_{\text{vis}}$ to force the newly learned visual features to be similar with the original ones. The UNIT model is updated to minimize the total loss:

$$\theta = \arg\min_{\theta} \mathcal{L}_{\text{lan}}(\{\mathcal{D}^I_{\times 1} \cup \mathcal{D}^T_{\times 4}\}) + \lambda \mathcal{L}_{\text{vis}}(\{\mathcal{D}^I_{\times 1}\}). \tag{5}$$

3.5 Inter-Scale Finetuning

After the intra-scale pretraining stage, UNIT is capable of processing both image and text recognition at their commonly used resolutions, namely, low-resolution images and high-resolution documents. However, we observed that this model lacks scale robustness when handling texts with larger fonts and images with larger dimensions, significantly limiting its generalization ability in various vision tasks. Naively add the scale-exchanged datasets in the intra-scale pretraining process may slow down the convergence of the model and making it prone to local optima. To address this, we further conduct inter-scale finetuning (see Fig. 2b), using high-resolution images and low-resolution documents.

Dataset Preparation. Here we build another scale exchanged dataset $\mathcal{D}^* = \mathcal{D}^I_{\times 4} \cup \mathcal{D}^T_{\times 1}$ with 1M samples for the inter-scale finetuning. We conduct a dataset $\mathcal{D}^I_{\times 4}$ comprising natural images annotated with detailed captions (each containing over 100 words). These images are resized to $\times 4$ times the original scale r, primarily drawn from the ShareGPT4V dataset [11], totaling 1M samples. Additionally, we create a dataset $\mathcal{D}^T_{\times 1}$ comprising 1M documents with an average font size $\times 4$ times larger than that of $\mathcal{D}^T_{\times 4}$. The document images are resized $\times r$ times, sourced from our synthetic dataset which includes cropped pages from digital books, advertisements, websites, and other sources containing short yet large font texts. We adopt the instruction prompt setting same to the intra-scale pretraining stage.

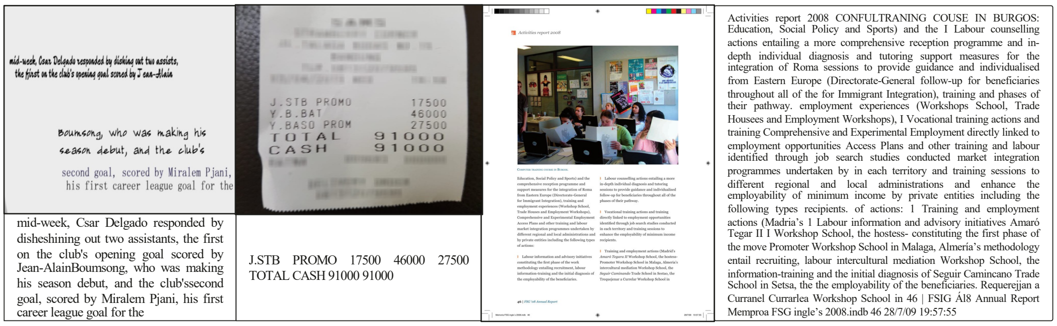
Random Feature Sampling. When dealing with natural images at higher resolutions, the resulting increase in visual tokens compared to lower resolutions can lead to gradient overflow during the training process, primarily due to the token-wise feature reconstruction loss. To address this challenge, we propose to randomly sample a set of visual tokens equal to the number in the low-resolution inputs and discarding other tokens.

Joint Optimization. This stage aims to enhance the scale-robust recognition ability for both image and text. We retain half of the data at the previous stage and then introduce the scale-exchanged data. The model is updated to minimize the total loss:

$$\theta = \arg\min_{\theta} \mathcal{L}_{\text{lan}}(\{\mathcal{D}^I_{\times 4} \cup \mathcal{D}^T_{\times 1} \cup \frac{1}{2}\mathcal{D}^I_{\times 1} \cup \frac{1}{2}\mathcal{D}^T_{\times 4}\}) + \lambda \mathcal{L}_{\text{vis}}(\{\frac{1}{2}\mathcal{D}^I_{\times 1} \cup \mathcal{D}^I_{\times 4}\}). \tag{6}$$

4 Experiments

Implementation Details. We select OpenCLIP ViT-H [43] with 32 layers and a hidden size of 1280 as our vision encoder. We choose OPT-125M [68] model with a hidden size of 768 as the language


mid-week, Csar Delgado responded by dishing out two assists,
the first on the club's opening goal scored by Jean-Alain
Boumsong, who was making his
season debut, and the club's
second goal, scored by Miralem Pjani,
his first career league goal for the
J.STB PROMO 17500
Y.B.BAT 46000
Y.BASO PROM 27500
TOTAL 91000
CASH 91000

mid-week, Csar Delgado responded by disheshining out two assistants, the first on the club's opening goal scored by Jean-AlainBoumsong, who was making his season debut, and the club'ssecond goal, scored by Miralem Pjani, his first career league goal for the

J.STB PROMO 17500 46000 27500 TOTAL CASH 91000 91000

Activities report 2008 CONFULTRANING COUSE IN BURGOS: Education, Social Policy and Sports) and the I Labour counselling actions entailing a more comprehensive reception programme and in-depth individual diagnosis and tutoring support measures for the integration of Roma sessions to provide guidance and individualised from Eastern Europe (Directorate-General follow-up for beneficiaries throughout all of the for Immigrant Integration), training and phases of their pathway. employment experiences (Workshops School, Trade Housees and Employment Workshops), I Vocational training actions and training Comprehensive and Experimental Employment directly linked to employment opportunities Access Plans and other training and labour identified through job search studies conducted market integration programmes undertaken by in each territory and training sessions to different regional and local administrations and enhance the employability of minimum income by private entities including the following types recipients. of actions: 1 Training and employment actions (Madria's 1 Labour information and advisory initiatives Amaró Tegar II I Workshop School, the hostess- constituting the first phase of the move Promoter Workshop School in Malaga, Almería's methodology entail recruiting, labour intercultural mediation Workshop School, the information-training and the initial diagnosis of Seguir Camincano Trade School in Setsa, the the employability of the beneficiaries. Requerejjan a Curranel Currarlea Workshop School in 46 | FSIG Ál8 Annual Report Memproa FSG ingle's 2008.indb 46 28/7/09 19:57:55

Figure 3: Visualization examples of text recognition. UNIT predicts accurate OCR results even across diverse scenarios, e.g., handwritten texts, receipts, and interleaved image-text documents. Please see clearly by zooming in. More promising examples are shown in the supplementary material.

Method	Backbone	#Params	Input	FUNSD	SROIE	CORD	SYN-L-val	MD-val
Donut [23]	Swin-B	260M	1280 × 960	9.08	8.94	16.64	44.78	5.07
Nougat [5]	SAM-ViT-B	247M	896 × 672	55.35	33.64	1.57	66.76	**86.71**
Vary* [53]	SAM-ViT-B	525M	1024×1024	21.01	9.84	12.89	91.20	59.30
RADIO* [44]	ViT-H/14	632M	896×896	26.12	10.42	10.01	93.90	37.57
UNIT (ours)	ViT-H/14	632M	896×896	**67.14**	**41.48**	**58.87**	**95.33**	78.50

Table 1: Comparison with ViT-based models for text recognition ability. The presented numbers are F1 values. An asterisk (*) indicates reimplemented results on our document datasets.

decoder in the two training stages. Our optimization strategy involves AdamW [32] with a weight decay 0.01. The initial learning rate to 5e-5, and changes with a cosine learning rate decay scheduler. The warmup ratio is set to 0.03, and the global batch size is 256. We set loss weights $\lambda = 2$ and $\mu = 0.2$. These settings are shared for both two training stages. Our model can be trained with any resolution below the maximum limit. We chose two primary resolutions for training to ensure consistent tensor sizes, simplifying batch processing and optimizing resource use. UNIT is trained on these two resolutions and tested on different ones. To ensure efficient data loading and processing during training, we preprocessed the document images by padding them to square shapes. During inference, we maintain the original aspect ratio of the input images.

4.1 Evaluation Benchmarks

Text Recognition: 1) Document-level OCR: We evaluate our model on three public OCR datasets: FUNSD (50 images), SROIE (347 images), and CORD (100 images), reporting the F1 score. Given the lack of large blocks of dense text in these datasets, we create two additional OCR evaluation datasets using English PDF files from arXiv. The first, SYN-L-val (200 images), consists of PDFs with small font sizes at approximately 1k resolution. The second, SYN-S-val (200 images), includes images cropped from these PDFs, resized to a lower resolution (e.g., 224), resulting in larger font sizes. **2) Markdown conversion:** The task requires structured text output from images of documents, capturing the stylistic and structural elements of the text in a markdown format. We add 1M markdown data (collected following Nougat [5]) into the inter-scale finetuning stage, endowing the ability of markdown conversion of our model. Due to the lack of cleaned markdown conversion validation set, MD-val, we conduct a dataset containing 82 images with markdown format annotations.

Image Recognition: 1) Zero-shot classification: We calculate the feature similarity between the [CLS] token of the visual encoder and the text features extracted from the corresponding CLIP text encoder. The evaluation is performed on the ImageNet-1K dataset, and top-1 accuracy is reported. **2) k-NN classification:** We first compute the [CLS] token of the visual encoder for the training images of ImageNet-1K, and then for each validation image, we utilize a weighted sum of the k nearest training vectors to select a label. **3) Semantic Segmentation:** we append a linear head onto the vision encoder and train it with the encoder frozen. The AdamW optimizer is used with a learning rate of 10^{-3}. The segmentation mIoU (%) is computed as in the MMSeg framework [13] on the ADE20k dataset. For both training and evaluation, we use an input size of 512×512.

Method	#Param.	ZS cls.	kNN cls.	Segm.
EfficientViT-L1 [7]	38M	71.73	79.90	33.12
SwinV2-S [29]	49M	74.70	81.12	35.57
ConvNext-B [31]	88M	75.43	81.73	38.95
MViTV2-B [27]	51M	75.92	81.39	41.39
NFNet-F3 [6]	254M	76.93	80.50	38.31
MaxViT-B [48]	119M	77.49	79.34	38.46
OpenCLIP-H/14 [43]	632M	77.19	81.10	40.04
RADIO-L/14 [44]	304M	77.25	84.03	48.70
E-RADIO-L/14 [44]	265M	77.87	83.73	45.50
RADIO-H/14 [44]	632M	78.62	84.17	49.01
UNIT (ours)	632M	**78.76**	**84.18**	**50.19**

Table 2: Comparison of image recognition capabilities with existing vision encoders.

Method	ZS cls.	OCR SYN-L-val		
		Prec.	Rec.	F1
w/o Image captioning				
$\lambda = 0$	N/A	93.03	90.63	91.81
$\lambda = 1$	74.20	92.18	89.43	90.78
$\lambda = 2$	76.24	94.13	91.72	92.91
with Image captioning				
$\lambda = 0$	N/A	95.17	93.39	94.26
$\lambda = 1$	75.16	92.56	89.26	90.88
$\lambda = 2$	**78.54**	**95.45**	**93.72**	**94.57**

Table 3: Ablation of the feature reconstruction loss weight λ with both the image captioning task during the intra-scale pretraining stage.

	Commonly-used Resolution				Exchanged Resolution			
Method	ZS cls.	OCR Prec	OCR Rec.	OCR F1	ZS cls.	OCR Prec	OCR Rec.	OCR F1
w/o inter-scale	78.54	95.45	93.72	94.57	0.40	51.79	28.53	34.99
w/o feat. sample	78.02	93.96	86.63	90.14	31.66	96.23	89.93	92.40
UNIT (ours)	**78.76**	**96.67**	**94.03**	**95.33**	**80.49**	**96.35**	**90.61**	**92.68**

Table 4: Ablation of the inter-scale finetuning stage and the random feature sampling strategy.

Downstream Vision Tasks: We replace the vision encoder in the LLava-1.5 [28] setting with our own encoder. We first pretrain the input embedding layers with the vision encoder and the LLM frozen, then conduct instruction tuning to finetune a Vicuna-7B model [42]. The learning rate of stage 3 is set to 5e-4 and a batch size of 512. We evaluate the trained models on general visual question answering (VQA) datasets including VQAv2 [17], GQA [20], and OKVQA [33], and also document comprehension datasets including ChartQA [34], DocQA [36], and InfoVQA [35].

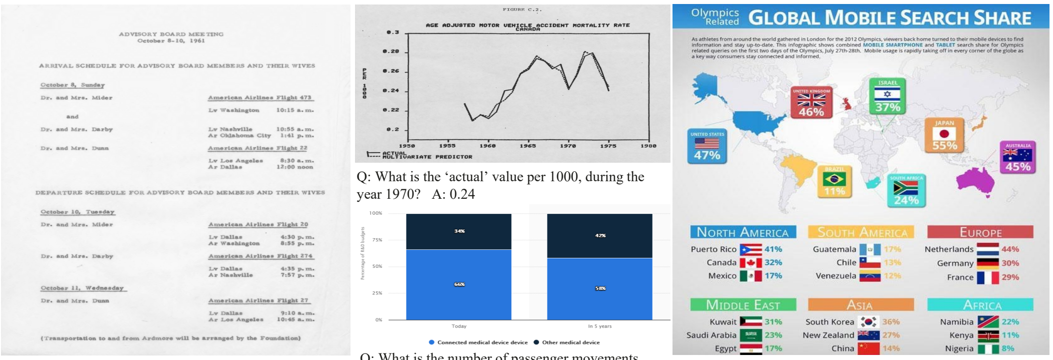
4.2 Comparison with Existing Approaches

Text Recognition. We compare UNIT with two ViT-based OCR expert model, including Donut [23] and Nougat [5], and two open-source vision encoders that support high-resolution inputs, including Vary [53] and RADIO [44]. We use the open-source weights of Donut-Base and Nougat for evaluation. As for Vary and RADIO, we retrain them on our dataset for a fair comparison. The models are evaluated on public OCR dataset like FUNSD [21], SROIE [19], CORD [19] and our synthetic datasets SYN-L-val and MD-val. As shown in Tab. 1, UNIT consistently outperforms these methods on OCR and markdown conversion tasks, except for Nougat, exhaustively trained on markdown data. This shows UNIT's superior ability to encode fine-detailed sequential and spatial information in documents, resulting in highly accurate OCR outcomes, see Fig. 3. Based on the fundamental text recognition ability, UNIT can be further finetuned with a small amount of data to adapt to highly specialized tasks, such as text recognition with grounding (in both natural images and pure texts), markdown conversion and Chinese OCR. Please refer to the supplementary for more details.

Image Recognition. To validate the image encoding ability of UNIT, we compare it with existing efficient vision encoders on zero-shot and kNN classification on ImageNet-1k, as well as a linear prob semantic segmentation benchmark. In Tab. 2, for fair comparision, all models share the same teacher knowledge, derived from DINOv2 and OpenCLIP-H. The compared models directly use these two models for training feature reconstruction, and UNIT employs one of the best-performing students, RADIO-H, as the teacher encoder. From the results shown in Tab. 2, we observe that UNIT achieves the best performance among all these models, demonstrating its ability to effectively preserve the fundamental image encoding capabilities.

4.3 Ablation Studies

About intra-scale pretraining. As shown in Table 3, removing the additional image captioning task described in Sec. 3.3 leads to a significant drop in performance on zero-shot classification (from 78.54% to 76.24%). This demonstrates that the image captioning task can further enhance image recognition capabilities. Additionally, we ablate the feature reconstruction loss weight λ, finding


ADVISORY BOARD MEETING
October 8-10, 1961
ARRIVAL SCHEDULE FOR ADVISORY BOARD MEMBERS AND THEIR WIVES
October 8, Sunday
Dr. and Mrs. Mider
American Airlines Flight 473
Lv Washington 10:15 a.m.
and
Dr. and Mrs. Darby
Lv Nashville 10:55 a.m.
Ar Oklahoma City 1:41 p.m.
Dr. and Mrs. Dunn
American Airlines Flight 22
Lv Los Angeles 8:30 a.m.
Ar Dallas 12:00 noon
DEPARTURE SCHEDULE FOR ADVISORY BOARD MEMBERS AND THEIR WIVES
October 10, Tuesday
Dr. and Mrs. Mider
American Airlines Flight 20
Lv Dallas 4:30 p.m.
Ar Washington 8:55 p.m.
Dr. and Mrs. Darby
American Airlines Flight 274
Lv Dallas 4:35 p.m.
Ar Nashville 7:57 p.m.
October 11, Wednesday
Dr. and Mrs. Dunn
American Airlines Flight 27
Lv Dallas 9:10 a.m.
Ar Los Angeles 10:45 a.m.
(Transportation to and from Ardmore will be arranged by the Foundation)
FIGURE C.2.
AGE ADJUSTED MOTOR VEHICLE ACCIDENT MORTALITY RATE
CANADA
PER 1000
0.3
0.28
0.26
0.24
0.22
0.2
1950
1955
1960
1965
1970
1975
1980
ACTUAL
MULTIVARIATE PREDICTOR
100%
75%
50%
25%
0%
Percentage of R&D budgets
34%
66%
42%
58%
Today
In 5 years
● Connected medical device device
● Other medical device
Olympics Related
GLOBAL MOBILE SEARCH SHARE
As athletes from around the world gathered in London for the 2012 Olympics, viewers back home turned to their mobile devices to find information and stay up-to-date. This infographic shows combined MOBILE SMARTPHONE and TABLET search share for Olympics related queries on the first two days of the Olympics, July 27th-28th. Mobile usage is rapidly taking off in every corner of the globe as a key way consumers stay connected and informed.
UNITED STATES 47%
UNITED KINGDOM 46%
ISRAEL 37%
JAPAN 55%
AUSTRALIA 45%
BRAZIL 11%
SOUTH AFRICA 24%
NORTH AMERICA
Puerto Rico 41%
Canada 32%
Mexico 17%
SOUTH AMERICA
Guatemala 17%
Chile 13%
Venezuela 12%
EUROPE
Netherlands 44%
Germany 30%
France 29%
MIDDLE EAST
Kuwait 31%
Saudi Arabia 23%
Egypt 17%
ASIA
South Korea 36%
New Zealand 27%
China 14%
AFRICA
Namibia 22%
Kenya 11%
Nigeria 8%

Q: When was 'advisory board meeting' scheduled?
A: October 8-10, 1961

Q: What is the 'actual' value per 1000, during the year 1970? A: 0.24

Q: What is the number of passenger movements (in thousands) at Dover ports in the UK?
A: 10901

Q: From which country in the Middle East did 23% of searches come? A: saudi arabia

Figure 4: Visualization examples of downstream document analysis tasks. UNIT accurately recognizes tiny words and digits, providing correct answers for document-related questions from users.

Benchmark	224 × 224	336 × 336	448 × 448	672 × 672	896 × 896
Zero-shot classification	78.8	81.3	81.7	81.3	80.5
OCR F1	92.7	93.7	93.2	93.9	95.3

Table 5: Ablation studies on different input resolutions during inference.

the best performance when $\lambda = 2$. When the feature reconstruction loss $\lambda = 0$, the zero-shot classification failed. This is reasonable because zero-shot classification evaluates the alignment between the vision and text encoders. If $\lambda = 0$, the vision encoder cannot ensure proper alignment with the text encoder of OpenCLIP-H, leading to the failure of zero-shot classification.

About inter-scale finetuning. As shown in Table 4, "w/o inter-scale" represents the model trained without inter-scale finetuning. Commonly-used Resolution indicates that the input sizes are at the common resolution of the evaluated tasks, e.g., 224×224 for zero-shot classification and 896×896 for document OCR (SYN-L-val). Exchanged Resolution indicates 896×896 for zero-shot classification and 224×224 for document OCR (SYN-S-val). We can see that after inter-scale finetuning, UNIT demonstrates enhanced scale robustness on the exchanged resolution for both image and documents. "w/o feat. sample" represents the model trained without random feature sampling mentioned in Sec. 3.5. This setting causes a dramatic performance drop on zero-shot classification on at exchanged resolution, showing the effectiveness of random feature sampling.

Different resolutions. Our model supports arbitrary input sizes within the maximum limit. As shown in Table 5, our model, trained on two primary resolutions, generalizes well and maintains consistent performance in both zero-shot classification and OCR tasks.

4.4 Downstream Task Performance

Naive resolution. Vision encoders act as a key component in Large Vision-Language Models (LVLMs) to provide rich and detailed visual representations that enable effective cross-modal understanding and reasoning. Following LLava-1.5 [28], we build LVLMs using UNIT or other vision encoders based on the Vicuna-7B LLM. The vision encoders are fixed during LVLM training. Images are padded and resized to one of two primary resolutions based on the task requirements. A lower resolution (e.g., 224x224) is used for global understanding tasks (e.g., VQA), while a higher resolution (e.g., 896x896) is used for tasks requiring fine-detail perception (e.g., DocQA). Each image is represented as 256 visual tokens through the input embedding layer. The LLM is finetuned with LoRA [18]. The multi-modal pretraining process uses 4M image-caption data (randomly extracted from Conceptual Caption [46] and the LAION-COCO [45]) and 200k document OCR data (randomly sampled from the training set). In the Supervised Finetuning (SFT) stage, we use the LLaVA-80k or LLaVA-CC665k along with the train set of DocVQA [36] and ChartQA [34] as the fine-tuning dataset. In Tab. 6, UNIT significantly outperforms the teacher model RADIO and other compared models on the document analysis tasks, including ChartQA, DocQA and InfoVQA, while maintains comparable performance on image understanding tasks. From Fig. 4 we can see that UNIT can extract the texts in documents, charts or websites, showing its potential in real-world document analysis applications.

Method	#Param	Document Analysis			Image Understanding		
		DocQA	ChartQA	InfoVQA	VQAv2	GQA	OKVQA
CLIP-L [43]	304M	22.3	23.6	25.0	72.8	61.2	55.8
DINOv2 [41]	632M	14.7	15.9	-	-	63.9	-
SAM-L [24]	632M	13.9	15.0	-	-	51.0	-
Vary [53]	525M	42.8	41.8	-	-	42.6	-
RADIO-H [44]	632M	18.2	20.2	24.5	72.9	61.8	55.9
UNIT (ours)	632M	**43.0**	**46.0**	**28.7**	72.3	60.4	55.5

Table 6: Comparison of the performance of various vision encoders with naive resolution.

Method	ChartQA	DocVQA	InfoVQA	OCRBench	GQA	OKVQA	MME	MathVista
CLIP-L [43]	52.0	57.2	29.3	382	62.3	57.0	1503.6	42.7
SigLIP [65]	56.5	62.0	29.7	429	63.0	61.1	1489.4	44.2
UNIT (ours)	**61.0**	**65.5**	**31.9**	**480**	**63.9**	**61.5**	**1529.8**	**44.6**

Table 7: Comparison of commonly used vision encoders in LVLMs with high-resolution grid slicing.

High resolution. In the naive resolution setting, each image is represented using 256 tokens. However, this may be insufficient for capturing the details of images with extensive content, such as densely organized documents containing thousands of words. To address this, we evaluate UNIT in Llava-Next with a grid slicing strategy and compare it with two commonly used vision encoders in LVLMs. This method divides each image into several grids based on their aspect ratio. For instance, with a maximum of 4 grids, possible configurations include $\{2 \times 2, 1 \times 2, 1 \times 3, 2 \times 1, 3 \times 1\}$, with each grid sized at $M \times M$. Here, M corresponds to the resolution used during model pretraining—336 for CLIP-L and 384 for SigLIP. And we set 448 for UNIT. After the vision encoder, we use C-Abstractor [37] as the input embedding layer. It generates 256 visual tokens for each grid, which are then concatenated to create the final image representation. During the multi-modal pretraining stage, the input embedding layer is initially trained on Llava665k. Following this, the LLM and the last few layers of the vision encoder are trained using 1.5M image captioning data. In the SFT stage, all parameters of the LVLM are fine-tuned using 1.2M SFT data. As shown in Tab. 7, our method significantly outperforms the compared models on document-oriented QA tasks and demonstrates comparable performance on other QA tasks. CLIP-L and SigLIP are initially pretrained only on natural images and subsequently fine-tuned with document images for downstream tasks. In contrast, UNIT is pretrained on both natural images and documents using fundamental recognition tasks such as image captioning and OCR. Our approach offers two key advantages: First, pretraining with fundamental recognition tasks enhances UNIT's versatility and suitability for downstream document tasks. Second, the early integration of text recognition capabilities in UNIT provides a robust initialization for LVLM training, leading to more stable training and faster convergence.

5 Conclusion

We propose UNIT, a simple yet effective framework that unifies image and text recognition in a single vision encoder, without increasing deployment or inference cost. UNIT incorporates a lightweight language decoder for text recognition and a lightweight vision decoder to preserve image recognition capabilities. Trained with multi-scale inputs, UNIT processes images and documents at their conventional resolutions during intra-scale retraining for basic recognition, and at exchanged resolutions during inter-scale finetuning to enhance scale robustness. Extensive experiments demonstrate that UNIT significantly outperforms document-specific models, maintains core image encoding ability, and benefits downstream document analysis tasks when integrated into LVLMs, showing great potential in processing interleaved text and image information in real-world scenarios.

References

[1] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] Alibaba. Introducing qwen-7b: Open foundation and human-aligned models (of the state-of-the-arts). `https://github.com/QwenLM/Qwen-7B`, 2023.

[3] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] D. Bautista and R. Atienza. Scene text recognition with permuted autoregressive sequence models. *arXiv preprint arXiv:2207.06966*, 2022. cs.

[5] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

[6] A. Brock, S. De, S. L. Smith, and K. Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning (ICML)*, pages 1059–1071. PMLR, 2021.

[7] H. Cai, J. Li, M. Hu, C. Gan, and S. Han. Efficientvit: Multi-scale linear attention for high-resolution dense prediction, 2023.

[8] Y. Cao, Z. Yihan, H. Xu, and D. Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020.

[10] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[11] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[12] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang. Text recognition in the wild: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–35, 2021.

[13] M. Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[14] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2988–2997, 2021.

[15] D. H. Diaz, S. Qin, R. Ingle, Y. Fujii, and A. Bissacco. Rethinking text line recognition models. *arXiv preprint arXiv:2104.07787*, 2021. cs.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Machine Learning (ICLR)*, 2022.

[19] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:72096–72109, 2023.

[20] D. A. Hudson and C. D. Manning. GQA: a new dataset for compositional question answering over real-world images. *CoRR*, abs/1902.09506, 2019.

[21] G. Jaume, H. K. Ekenel, and J.-P. Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.

[22] D. Kim, A. Angelova, and W. Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11144–11154, 2023.

[23] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, pages 498–517. Springer, 2022.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[25] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR, 2023.

[26] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.

[27] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, 2022.

[28] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

[29] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[31] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s, 2022.

[32] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Machine Learning (ICLR)*, 2019.

[33] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.

[34] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[35] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[36] M. Mathew, D. Karatzas, and C. 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.

[37] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

[38] I. Misra, R. Girdhar, and A. Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2917, 2021.

[39] B. Moysset, C. Kermorvant, and C. Wolf. Full-page text recognition: Learning where to start and when to stop. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 871–876. IEEE, 2017.

[40] OpenAI. Chatgpt. `https://openai.com/blog/chatgpt/`, 2023.

[41] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[42] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[44] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12490–12500, June 2024.

[45] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[46] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565, 2018.

[47] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

[48] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li. Maxvit: Multi-axis vision transformer, 2022.

[49] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1457–1464. IEEE, 2011.

[50] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 3304–3308. IEEE, 2012.

[51] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, 2021.

[52] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23034–23044, 2023.

[53] H. Wei, L. Kong, J. Chen, L. Zhao, Z. Ge, J. Yang, J. Sun, C. Han, and X. Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.

[54] H. Wei, L. Kong, J. Chen, L. Zhao, Z. Ge, E. Yu, J. Sun, C. Han, and X. Zhang. Small language model meets with reinforced vision vocabulary. *arXiv preprint arXiv:2401.12503*, 2024.

[55] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[56] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, 2021.

[57] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:12077–12090, 2021.

[58] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18134–18144, 2022.

[59] R. Yang, L. Song, Y. Li, S. Zhao, Y. Ge, X. Li, and Y. Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*, 2023.

[60] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, Y. Dan, C. Zhao, G. Xu, C. Li, J. Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

[61] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, G. Xu, C. Li, J. Tian, Q. Qian, J. Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

[62] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[63] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051, 2024.

[64] Y. Zeng, C. Jiang, J. Mao, J. Han, C. Ye, Q. Huang, D.-Y. Yeung, Z. Yang, X. Liang, and H. Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15244–15253, 2023.

[65] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023.

[66] Q. Zhang, J. Zhang, Y. Xu, and D. Tao. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.

[67] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8552–8562, 2022.

[68] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[69] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *URL https://arxiv. org/abs/2205.01068*, 3:19–0, 2023.

[70] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[71] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10323–10333, 2023.

A Appendix / supplemental material

A.1 Computer Resources

We implement our method using PyTorch 2.1.0 with CUDA 11.7. The training process requires 128 Ascend 910B GPUs (each with 64 GB of memory). Each training stage runs for one epoch, totaling 150 hours of training time.

A.2 Limitations

Integrating text and image recognition in a single model may entail trade-offs, potentially yielding slightly inferior performance in highly specialized tasks compared to models exclusively dedicated to text or image recognition. For example, as illustrated in the Table 1 of our paper, UNIT's performance in the markdown conversion task (a highly specialized OCR task) falls short of Nougat, a model exhaustively trained on markdown data and specifically optimized for markdown conversion.

A.3 Broader Impacts

Positive Societal Impacts: 1) Enhanced Accessibility: By unifying image and text recognition within a single model, UNIT could potentially improve accessibility for individuals with visual impairments. The model's ability to accurately recognize and interpret both images and text could facilitate the development of assistive technologies such as screen readers and optical character recognition (OCR) systems. 2) Efficiency in Document Analysis: UNIT's capability to seamlessly handle both image and text recognition tasks could streamline document analysis processes in various domains, including healthcare, education, and legal sectors. This could lead to increased efficiency and productivity in tasks such as document digitization, content extraction, and information retrieval. 3) Technological Advancement: The development of novel training frameworks like UNIT contributes to the advancement of computer vision and natural language processing technologies. Such advancements have the potential to drive innovation in diverse fields, including artificial intelligence, robotics, and human-computer interaction.

Possible Negative Societal Impacts: 1) Privacy Concerns: The improved accuracy and efficiency of text recognition enabled by models like UNIT may raise concerns regarding privacy and data security. Increased capability to extract and analyze text from images could potentially infringe upon individuals' privacy rights if not properly regulated or safeguarded. 2) Job Displacement: The automation of document analysis tasks facilitated by UNIT and similar models could lead to job displacement in industries that rely heavily on manual data entry and document processing. While this could result in increased efficiency and cost savings for businesses, it may also contribute to unemployment and economic inequality.

NeurIPS 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: We believe that the main claims reflect the areas we are targeting: vision fundation models and large-scale vision-language models.

   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: Please see the Appendix.

   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: Our paper does not include theoretical results.

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: See the Section 3 Methodology and Section 4 Experiments.

   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**

   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: [No]

Justification: According to the code and data authorization rule of our institution, we are not allowed to directly attach the code and data in the anonymous submission. We will apply for a open-source license once our paper is accepted.

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: See the Section 4 Experiments for implementation 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: Our experiments are quite stable with multiple runs. All results are observed via a fixed seed.

   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: See Appendix for more details.

   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: The research is conducted with the NeurIPS code of Ethics in every respect.

   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: See Appendix.

    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.
    - 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: Our 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: We have cited the creators of all existing assets. All the used existing assets are open-source.

    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: [NA]

    Justification: The paper does not release new assets.

    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.
    - 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.