On the Design Fundamentals of Pixel Text Representation Learning

Anonymous ACL submission

Abstract

Text-rich visual inputs require models that can read, retrieve, and compress language directly in pixel space, yet existing pixel-text encoders struggle with fixed resolution pretraining, visual shortcut learning, weak visual grounding, and multilingual visual text understanding. In this work, we investigate the fundamental design principles required for robust visual text representation learning. Through systematic ablations, we identify four critical components: variable image resolutions and rendered font sizes provide spatial proxies for high-resolution document generalization; natural image–text pairs are indispensable for grounding and prevent text-only collapse; layout-aware rendering helps prevent pixel-level shortcuts; and a two-stage multilingual curriculum enables effective cross-lingual alignment. By integrating these principles into a scalable training recipe, we train PIXEL LINGUIST II, a native-resolution vision encoder trained with on-the-fly rendering, unified contrastive grounding, and a multilingual curriculum over 280M training examples. PIXEL LINGUIST II sets new state-of-the-art results on English, cross-lingual, and multilingual Visual STS and ViDoRe, while also enabling better MLLM downstream evaluation. Notably, PIXEL LINGUIST II remains robust under 80% visual token compression, showing great promise for optical context compression.

1 Introduction

Vision-language representation learning has become central to cross-modal retrieval and retrieval-augmented generation (RAG). While dual-encoder models excel on natural images and short captions, they are less suited to text-rich visual inputs, such as documents, infographics, and charts, where retrieval requires fine-grained reading, layout understanding, and document-level semantics.

Pixel-based text representation learning offers a unified alternative: render text directly as RGB images and encode both natural images and rendered text with a single vision encoder. Prior work has progressively shown that ViT encoders can learn language representations from pixel inputs (Rust et al., 2022), that contrastive learning improves their discriminability (Tschannen et al., 2023), and that scaled rendered-text training enables visual, topical, reasoning, and cross-lingual alignment (Xiao et al., 2024). Despite this progress, robust pixel-based text representation learning still hinges on four coupled challenges: **resolution mismatch**, **visual shortcut learning**, **multimodal grounding**, and **multilingual visual text perception**. These axes determine whether a visual text encoder can move beyond synthetic rendered snippets to real-world document understanding.

Instead of simply scaling data and parameters, we ask: *What are the essential design principles required to learn generalized visual text representations?* Our controlled ablations answer this question through four research questions:

RQ1: How can computationally efficient low-resolution pretraining generalize to high-resolution documents at inference? We find that variable natural-image resolutions and rendered font sizes act as spatial proxies, allowing small-canvas pretraining to extrapolate to dense, high-resolution documents.

RQ2: What role does multimodal grounding play in visual text representation learning? Natural image–text pairs remain necessary even for text-centric targets: removing them causes severe dense document retrieval degradation, while joint training grounds text semantics in real-world visual contexts.

RQ3: How does layout diversity in text rendering affect representation quality? Fixed fonts and plain canvases trigger pixel-level shortcut learning and near-collapse in document retrieval, showing that text rendering with diverse layouts is critical for semantic transfer.

RQ4: What training curriculum is required for multilingual pixel-space semantics? We find that a two-stage curriculum works best: large-scale unsupervised multilingual pretraining builds foundational capability for multilingual visual text perception, which is then activated and aligned across languages through semantic mid-training.

Building on these findings, we scale the design principles into a concrete training recipe for pixel-based text representation learning. We propose PIXEL LINGUIST II, a unified pixel-based vision-language representation framework for robust understanding of text in the visual modality. As illustrated in Figure 1, PIXEL LINGUIST II combines four components:

1. **Layout-aware visual augmentation** renders text on the fly with diverse fonts, backgrounds, spatial arrangements, and visual perturbations, encouraging the model to encode semantics rather than memorizing superficial appearances.
2. **Native-resolution encoding** adopts a Native-resolution Vision Transformer (NaViT) architecture (Dehghani et al., 2023; Bai et al., 2025) supports variable image resolutions and aspect ratios. This design, when combined with our data pre-processing engine, enables the model to learn robust semantic extrapolation to extremely high-resolution inputs at test time.
3. **Unified contrastive grounding** jointly trains on natural image–text pairs and rendered text–text pairs under a single contrastive objective, learning text semantics in the visual modality while grounding them in real-world visual concepts.
4. **A multilingual training curriculum** scales learning to 280M examples through a two-stage pipeline: massive unsupervised multilingual visual text pretraining followed by high-quality semantic mid-training.

Extensive experiments validate both the design analysis and the resulting model. PIXEL LINGUIST II sets new state-of-the-art results across English, cross-lingual, and multilingual Visual STS benchmarks, achieves strong performance on the challenging ViDoRe visual document retrieval benchmark, and improves downstream performance when used as the vision encoder in multimodal large language models. Notably, visual text representations of PIXEL LINGUIST II remain robust even when up to 80% of visual tokens are compressed,

2 Design Fundamentals of Pixel Text Representation Learning

While unified visual text encoding offers a highly elegant architecture, current vision encoders remain constrained by fundamental bottlenecks: inflexible fixed-resolution processing, a lack of real-world multimodal grounding, and severe sensitivity to visual appearances.

To overcome these limitations, we explore essential design principles required to learn generalized, real-world visual text representations. Before running a massive-scale pretraining (Section 3), we first devise a number of controlled pretraining experiments using a compact 13M-example ablation dataset consisting of natural images and rendered text, resulting in four deisgn fundamentals for generalized visual text representation learning.

2.1 The Resolution Paradox and Spatial Proxies

A central challenge in pixel-based text encoding is the discrepancy between training and inference resolutions. Processing dense, high-resolution documents (e.g., 4K PDFs) requires encoding massive amounts of spatial information, yet pretraining is typically constrained to smaller, fixed-size canvases (e.g., 224×224 pixels) for computational efficiency. This contradiction prompts our first inquiry:

RQ1: How can a fixed-resolution canvas in training generalize to high-resolution documents in inference?

We hypothesize that computationally prohibitive high-resolution pretraining might not be strictly necessary if the network can learn the underlying concept of spatial scale through alternative means. To test this, we explore whether two factors—*resolutions in natural images* and *variable font sizes in rendered text*—can act as effective "spatial proxies" that enable the model to generalize to high-resolution documents without directly training on them. We ablate these variables during pretraining and evaluate the resulting models on resolution-sensitive document retrieval tasks (summarized in the top section of Table 1).

In our controlled setup, we enforce computational efficiency while preserving variance: (1) we dynamically resize the longest side of all natural images to 224 pixels, allowing the shortest side to vary and thus preserving aspect ratio diversity without inflating compute; and (2) we render tex-

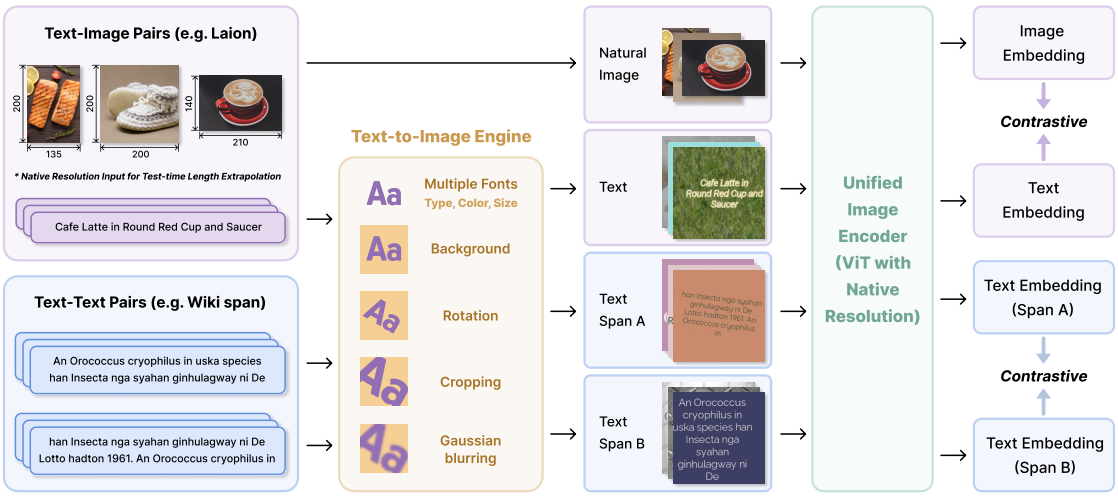

Figure 1: Overview of PIXEL LINGUIST II. We construct two types of training data: natural image–text pairs and text–text pairs. Textual inputs are rendered on-the-fly into images with diverse layouts using a text-to-image rendering engine. A native-resolution ViT encodes both natural images and rendered visual text within a unified pixel space. The model is trained by a contrastive learning objective.

Pretraining Setting	Data Type	Examples	Batch Size	ArxivQA	InfoVQA	TabFQuad	TatDQA	Average
Spatial Proxies (See Sec. 2.1)								
Full Components	Nat. Image + Ren. Text	13M	12288	14.99	56.69	56.05	23.60	**37.83**
Fixed Image Size (224 × 224)	Nat. Image + Ren. Text	13M	12288	9.61	46.33	55.23	22.61	33.45
Fixed Font Size (16)	Nat. Image + Ren. Text	13M	12288	7.83	48.28	47.89	19.86	30.97
Multimodal Grounding & Shortcut Removal (See Sec. 2.2 & 2.3)								
Fixed Font + Plain Canvas	Rendered Text Only	7M	12288	0.59	0.61	4.20	1.18	1.65
Full Scaled-up Run (See Sec. 3)								
Full Components (Ours)	Nat. Image + Ren. Text	104M	32768	35.81	67.61	63.96	27.90	**48.82**

Table 1: Deconstructing the Pixel Pretraining Pipeline. We evaluate spatial proxies and multimodal grounding using controlled, small-scale ablations on resolution-sensitive ViDoRe tasks. Performance is measured in nDCG@5.

tual inputs onto a fixed 224 × 224 canvas while randomly sampling font sizes between 12 and 22.

Our observations reveal a clear trend: when we remove variance in natural image dimensions by resizing them to a static 224 × 224 resolution, average retrieval performance drops from 37.83 to 33.45. More critically, when we fix the rendered text to a uniform font size, performance degrades further to 30.97. Varying font sizes on a small canvas forces the model to encode textual features at multiple spatial frequencies. This confirms that these variances implicitly equip the model to generalize to high-resolution, dense documents during inference, completely bypassing the need to pretrain on massive synthetic document canvases.

2.2 The Necessity of Multimodal Grounding

A persistent open question in pixel-based representation learning is whether natural images are actually required if the target domain is primarily text. This leads to our second research question:

RQ2: Is synthetic rendered text alone sufficient for real-world visual text understanding? What is the role of multimodal grounding?

To answer this, we explore the boundaries of a purely synthetic visual space. We train a variant exclusively on rendered text pairs, completely removing natural image–text pairs from the pretraining corpus, mirroring approaches in previous work (Rust et al., 2022; Xiao et al., 2024).

We observe that while this text-only variant maintains competitive performance on simple semantic matching tasks like Visual STS, it suffers a substantial performance drop on the more complex ViDoRe benchmark. Furthermore, when we completely isolate the model by stripping away both natural images and the spatial proxies validated in Section 2.1 (i.e., training purely on rendered text with a fixed font size and static plain canvas), the degradation becomes catastrophic. As shown in the middle section of Table 1, this text-only fixed template setting yields a near-zero average document

retrieval score of just 1.65.

These findings suggest that training an encoder in an isolated, synthetic pixel space does not offer a valid solution for real-world document understanding. Natural image–text pairs act as a fundamental regularizer that grounds synthetic textual semantics in real-world visual contexts and exposes the model to the heterogeneous layout structures necessary for document processing. Consequently, joint multimodal training is strictly required to prevent representation collapse.

Later in Figure 3, we conduct the same ablation using our full-scale data, showing similar findings. This suggests that the necessity of multimodal grounding can not be bypassed by data scale alone.

2.3 Mitigating Shortcut Learning via Layout Augmentation

A related bottleneck in pixel-based text encoding is shortcut memorization. When text is rendered using fixed visual templates (e.g., uniform fonts and plain backgrounds), vision encoders naturally gravitate toward overfitting to superficial visual attributes. Returning to the theme of visual variation established in RQ1, we ask:

RQ3: How does layout diversity affect representation quality?

The vulnerability caused by this sensitivity is evident in our core ablations (Table 1). As previously noted, stripping away layout diversity by fixing the font size significantly degrades retrieval performance from 37.83 to 30.97. Furthermore, removing both font variance and background diversity (the "Plain Canvas" setting) leads to the catastrophic collapse observed in Section 2.2.

These ablations show that models must be forced to abstract away from pixel-level shortcuts. Thus, we introduce layout-aware visual augmentation. During on-the-fly text rendering, we inject structured layout diversity on fonts and backgrounds, detailed in Section 3.2. By ensuring the model never encounters the exact same visual instantiation of a text twice, we effectively force the encoder to prioritize semantic structure over visual shortcuts.

2.4 The Data Curriculum: Activating Multilingual Pixels

Having established the fundamental rendering and grounding principles, our final exploration focuses on the training trajectory itself. While high-quality curated semantic pairs are sufficient to train an English-only visual text encoder, extending this capability to a global multilingual pixel space introduces a severe bottleneck. Character sets such as Arabic, Chinese, and Korean exhibit vastly different visual and spatial structures compared to the Latin alphabet. Thus, we ask:

Curriculum Setting	Cross-lingual	Multilingual
Curated Semantic Pairs Only	53.83	61.79
Unsup. Pretraining + Curated Pairs	**57.16**	**65.27**

Table 2: Data Curriculum Ablation on Visual STS (Spearman Correlation).

RQ4: What curriculum is required to inject foundation capabilities for multilingual visual text understanding?

We investigate whether a model can learn cross-lingual semantic alignment purely from high-quality curated pairs, or if it fundamentally requires prior perceptual knowledge of these scripts. We evaluate two curriculum settings on the cross-lingual and multilingual subsets of Visual STS: one model trained *exclusively* on high-quality semantic pairs (Mid-Train Only), and another that first undergoes large-scale unsupervised contrastive pretraining on highly dense multilingual Wikipedia articles before semantic tuning (Pretrain + Mid-Train).

As shown in Table 2, relying solely on curated cross-lingual semantic pairs yields a performance ceiling. However, introducing a foundational stage of unsupervised pretraining provides a consistent boost of $\sim$3.3 to 3.5 absolute points across diverse languages. This establishes a core data curriculum principle: massive unsupervised multilingual visual text pretraining acts as a foundational training phase to inject multilingual visual text perceptual capability, which is subsequently *activated* and refined during the semantic mid-training phase.

3 Instantiating the Recipe: PIXEL LINGUIST II

Having established the fundamental design principles for optical text representation, we instantiate our methodology at scale. Our resulting model, PIXEL LINGUIST II, integrates native-resolution processing, multimodal grounding, layout-aware augmentation, and a strict data curriculum into a unified vision-only architecture.

3.1 Native-Resolution Encoding Architecture

To fully leverage the spatial proxies identified in Section 2.1 (i.e., variable image resolutions and

font sizes), the vision backbone must natively support arbitrary aspect ratios and resolutions without lossy resizing. We adopt a Native-resolution Vision Transformer (NaViT) architecture, initializing the ViT parameters from Qwen2.5-VL's ViT. By processing a variable number of visual tokens rather than relying on fixed-grid interpolation, the encoder preserves the fine-grained structural integrity of dense document layouts and small text. Additionally, a 2×2 pooling layer is applied to compress adjacent visual tokens, balancing semantic capability and encoding efficiency during modeling.

3.2 Layout-Aware Rendering Engine

To implement the augmentation requirements established in Section 2.3, we develop an on-the-fly text-to-image rendering engine. Textual inputs are rendered dynamically at each epoch, ensuring the model never ground text semantics in visual shortcuts. We sample from 393 unique fonts (Table 6) across languages, and stochastically apply background variations, including brightness jittering, Gaussian blur, and over 5,000 distinct textured backgrounds from the Describable Textures Dataset (DTD) (Cimpoi et al., 2014).

3.3 Unified Contrastive Grounding

Based on the multimodal grounding requirements established in Section 2.2, we design a data recipe for unified contrastive grounding, incorporating both text-text pairs and text-image pairs. For **text-text pairs**, we leverage both high-quality multilingual text pretraining copus (used in first-stage training) and high-quality multilingual text pair datasets (used in second-stage training), detailed in the next subsection (Section 3.4). For **image-text pairs** we sample 26M natural image-text pairs sourced from LAION-2B (Schuhmann et al., 2022) to maintain real-world grounding. This design serves as a regularizer to maintain the model's world knowledge, preventing the model from learning textual semantics purely from shapes.

3.4 The Scaled Multilingual Curriculum

To faciliate high multilingual visual text understanding capability, we instanstiate the two-stage training recipe established in Section 2.4 consisting of two text dataset types: (1) multilingual pretraining corpus. (2) high-quality text pairs.

For **multilingual pretraining corpus** (referred to as *Text Corpus 1*), we leverage multilingual Wikipedia pretraining corpus of 62M documents. For each document, we randomly crop 25% to 50% each document twice to serve as unsupervised positive pairs (Izacard et al., 2021). For **high-quality semantic text pairs** (referred to as *Text Corpus 2*), we curate 26M pairs from high-quality datasets used for text embedding model training.

Combining our multimodal and multilingual dataset recipes, the training curriculum is divided into two phases:

- **Stage 1: Foundational Pretraining** combines *Text Corpus 1* (62M examples) and *image-text pairs* (26M examples)
- **Stage 2: Semantic Mid-Training:** combines *Text Corpus 2* (26M examples) and *image-text pairs* (26M examples)

Each stage is run for 2 epochs, resulting in a total examples seen of 280 millions. Training implementation details are in Appendix A.

4 Main Results

Overview We evaluate PIXEL LINGUIST II on Visual Semantic Textual Similarity (Visual STS) and Visual Document Retrieval (VDR), comparing against 18 competitive baselines, including CLIP (Radford et al., 2021), OpenCLIP (Ilharco et al., 2021), DataComp-CLIP (Gadre et al., 2023), SigLIP (Zhai et al., 2023), and EVA-CLIP (Sun et al., 2023). Tables 3 and 4 compare strongest baselines on Visual STS and VDR, respectively (See Tables 9 and 10 for all model results).

To further stress-test visual text understanding beyond English, we evaluate PIXEL LINGUIST II on the cross-lingual and multilingual subsets of Visual STS. Cross-lingual results are summarized in Table 5 (full results in Table 11), while multilingual results are reported in Table 12 (Appendix F).

Visual Semantic Textual Similarity (English) As shown in Table 3, the variant of PIXEL LINGUIST II pretrained only on the *Mid-Training* datasets already achieves state-of-the-art performance across all English Visual STS tasks. Notably, it outperforms the largest existing vision encoders despite being ∼1/7 in model parameters and trained on approximately ∼1/87 examples seen.

Applying standard AllNLI fine-tuning with ∼270K examples further improves performance, yielding an additional ∼5-point gain in Spearman correlation. The strong performance of PIXEL LINGUIST II on Visual STS **demonstrates its ability to capture semantic textual similarity directly**

Model name	v-STS12	v-STS13	v-STS14	v-STS15	v-STS16	v-STS17	v-STS-b	Avg.
google/siglip-base-patch16-224	63.19	55.40	57.99	73.07	67.79	77.78	54.50	64.25
openai/clip-vit-large-patch14	53.89	66.78	55.98	72.03	70.49	75.26	56.74	64.45
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	57.00	62.25	58.62	74.40	70.57	76.69	58.99	65.50
openai/clip-vit-base-patch16	63.82	63.26	56.99	73.32	68.91	78.18	57.93	66.06
google/siglip-so400m-patch14-384	61.90	62.95	60.58	76.17	73.48	78.41	62.63	68.02
EVA02-CLIP-bigE-14	62.24	62.36	62.17	77.41	73.63	80.96	62.85	68.80
google/siglip-large-patch16-384	66.30	62.08	61.66	77.11	73.27	79.58	66.59	69.51
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	62.81	68.16	65.50	78.67	74.89	79.97	66.54	70.93
EVA02-CLIP-bigE-14-plus	63.36	68.00	66.38	79.45	75.26	82.87	68.59	71.99
			Backbone					
Qwen2.5-VIT	47.50	36.49	30.95	54.69	53.71	63.87	38.63	46.55
			Ours					
PIXEL LINGUIST II (mid-training only)	65.78	70.00	67.76	82.39	76.99	84.83	75.30	74.72
PIXEL LINGUIST II (mid-training + finetuning)	76.60	75.94	75.07	85.17	79.65	85.25	80.93	79.80

Table 3: PIXEL LINGUIST II Performance on Visual STS Tasks (English-only) (Xiao et al., 2024, 2025), which renders traditional STS tasks in NLP as image-only tasks. This task assesses the ability of vision encoders on textual semantic understanding on text-rich images.

Model name	DocVQA	InfoVQA	Sft Proj.	AI	Tabfquad	Tatdqa	Avg.
			Baselines				
openai/clip-vit-base-patch16	14.60	51.70	7.13	22.86	17.61	4.71	19.77
google/siglip-base-patch16-224	16.04	46.11	3.71	25.27	29.04	5.08	20.87
EVA02-CLIP-bigE-14	16.35	54.80	10.14	33.53	28.80	7.09	25.12
EVA02-CLIP-bigE-14-plus	16.84	55.19	12.76	34.57	30.36	7.52	26.21
openai/clip-vit-large-patch14	16.69	62.44	17.05	38.25	30.95	11.00	29.40
laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K	19.68	55.61	16.19	47.20	30.70	15.27	30.78
google/siglip-large-patch16-256	22.39	54.09	9.13	43.40	49.81	12.38	31.87
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	20.44	60.90	25.02	55.42	35.02	16.21	35.50
google/siglip-so400m-patch14-384	31.28	69.73	25.04	67.78	60.29	27.52	46.94
			Ablation				
google/siglip-so400m-patch14-384	31.28	69.73	25.04	67.78	60.29	27.52	46.94
google/siglip-so400m-patch14-384 (vision-only eval.)	12.05	34.70	8.96	34.54	34.67	11.81	22.79
Δ Performance Difference	19.23↓	35.03↓	16.08↓	33.24↓	25.62↓	15.71↓	24.15↓
			Backbone				
Qwen2.5-VIT	0.94	0.74	0.93	0.89	7.15	2.2	2.14
			Ours				
PIXEL LINGUIST II (mid-training only)	20.46	67.61	37.20	75.09	63.96	27.90	48.70
PIXEL LINGUIST II (mid-training + finetuned)	20.91	69.37	41.60	72.94	71.46	29.36	50.94

Table 4: Encoder performance of PIXEL LINGUIST II on Visual Document Retrieval (VDR) Tasks using ViDoRe subsets (Faysse et al., 2025) in MIEB-lite benchmark, compared with SOTA baseline encoder models.

from pixel inputs, highlighting effective zero-shot semantic understanding of rendered text.

Visual Document Retrieval (VDR) Unlike prior VDR-focused models, PIXEL LINGUIST II is primarily pretrained on synthetic visual text and does not explicitly train on real-world PDFs with complex document layouts. As such, its generalization to document retrieval tasks provides a stringent test of its visual text understanding capability.

Table 4 reports performance of ViDoRe subsets (Faysse et al., 2025) in MIEB-lite (Xiao et al., 2025). Overall, PIXEL LINGUIST II achieves state-of-the-art performance. In particular, it **exhibits strong capability in understanding *tables and charts*, and documents where *dense text is interleaved with structured visual elements***, resulting in gains of ∼5–12 nDCG@5 on `AI` and `TabFQuAD`, and a substantial improvement of 16.6 nDCG@5 on `ShiftProject` over previous SOTA.

Importantly, PIXEL LINGUIST II attains these results using a **vision-encoder-only** setup, in which textual queries are rendered as images and processed uniformly with visual documents. This setting places PIXEL LINGUIST II at an inherent disadvantage relative to CLIP-style dual-encoder models, which benefit from a dedicated text encoder that provides semantically rich textual embeddings.

To quantify this gap, we conduct a fair comparison against `siglip-so400m-patch14-384`, the

Model name	ko-ko	ar-ar	en-ar	en-de	en-tr	es-en	es-es	fr-en	it-en	nl-en	Avg.
EVA02-CLIP-bigE-14-plus	11.36	31.51	10.71	24.33	-10.05	20.18	59.20	36.12	28.60	33.18	24.52
EVA02-CLIP-bigE-14	10.97	29.99	13.49	22.76	6.39	29.03	57.16	36.66	33.43	26.16	26.60
google/siglip-base-patch16-224	21.00	25.03	14.36	31.20	24.80	29.32	69.85	35.70	27.46	28.98	30.77
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	14.38	32.39	12.21	36.74	14.99	30.44	69.77	39.77	36.44	34.83	32.20
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	19.39	33.39	19.49	43.78	16.68	27.99	62.58	39.32	28.59	37.33	32.85
openai/clip-vit-base-patch16	10.54	36.25	13.13	41.57	35.42	24.63	62.95	38.72	31.40	38.63	33.32
laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K	14.28	36.47	12.75	43.10	19.70	37.37	71.62	36.88	30.78	30.76	33.37
openai/clip-vit-large-patch14	11.07	39.12	18.95	45.71	39.70	36.76	70.11	44.06	40.17	41.63	38.73
google/siglip-so400m-patch14-384	13.65	45.76	11.22	46.07	30.62	40.08	73.62	46.36	36.45	44.95	38.88
Backbone											
Qwen2.5-ViT	51.34	52.45	22.07	24.77	22.58	16.71	65.44	32.05	26.00	26.43	33.98
Ours											
PIXEL LINGUIST II (mid-training)	51.13	50.96	2.09	66.00	54.59	55.51	70.66	63.00	61.91	62.46	53.83
PIXEL LINGUIST II (pre-training + mid-training)	57.51	50.13	1.44	67.4	55.86	61.42	75.82	67.73	67.88	66.36	57.16

Table 5: Encoder Performance of PIXEL LINGUIST II on Visual STS Tasks (Cross-lingual Tasks)

strongest SigLIP variant on ViDoRe. Shown in the middle section of Table 4, enforcing a unified visual processing pipeline for SigLIP leads to a substantial drop of 24.2 in nDCG@5 on average, underscoring the difficulty of performing VDR in a vision-only formulation and the robustness of PIXEL LINGUIST II in visual text understanding.

Visual STS (Cross-lingual and Multilingual) Beyond English, PIXEL LINGUIST II demonstrates strong and consistent capability in understanding multilingual text rendered as images. We evaluate this using the cross-lingual and multilingual Visual STS subsets derived from STS17 and STS-B, covering 11 languages. These include high-resource languages such as German, French, and Italian, as well as languages where prior vision encoders typically achieve near chance-level performance, such as Chinese, Russian, Korean, and Turkish.

Cross-lingual results comparing representative models are reported in Table 5 (More models in Table 11) and multilingual results in Table 12. Even when trained solely on the *Mid-Training* dataset, PIXEL LINGUIST II achieves SOTA performance, **outperforming the strongest SigLIP variant by ∼15% in Spearman correlation on cross-lingual tasks (Table 5) and over 16% on multilingual tasks (Table 12)**.

Moreover, combining **pretraining + mid-training** yields consistently better performance than mid-training alone. This indicates **the importance of foundational multilingual knowledge through pretraining on massive unsupervised rendered corpus, which is then effectively activated and enhanced through mid-training**, bringing multilingual visual text understanding closer to parity with English performance.

Evaluation on Downstream Tasks Described in Section 3.2, we evaluate PIXEL LINGUIST II on MLLM downstream tasks. Specifically, we compare with Qwen2.5-ViT (Bai et al., 2025) initialized from Qwen2.5-VL-7B-Instruct, by paring both with the same LLM (Qwen2.5-7B-Instruct) and conducting LLaVA-style post-training. Our model achieves a 2.75% average relative improvement over Qwen2.5-ViT across downstream tasks, validating its competence as a generalist vision encoder that provides comprehensive information to LLMs. Detailed results on these benchmarks are in Table 8 in Appendix D.2.

5 In-depth Analysis

Optical Context Compression Recent MLLMs utilize the visual modality as an efficient compression medium to alleviate textual context-length constraints (Wei et al., 2025; Cheng et al., 2025). However, aggressive visual token downsampling risks severe semantic loss. To evaluate the semantic density of PIXEL LINGUIST II, we downsample input images to induce varying compression rates across 32 Visual STS and ViDoRe tasks.

As shown in Figure 2, our model retains strong semantic representation even under substantial compression. On Visual STS, PIXEL LINGUIST II maintains performance parity with CLIP even when 60% of the visual tokens are discarded (retaining only 118 of the original 196 tokens). This efficiency is even more pronounced on dense documents: on the ViDoRe benchmark, PIXEL LINGUIST II continues to outperform the uncompressed CLIP baseline even at 80% token compression. These results confirm that our training recipe inherently yields highly compact representations, enabling extreme optical context compression with-

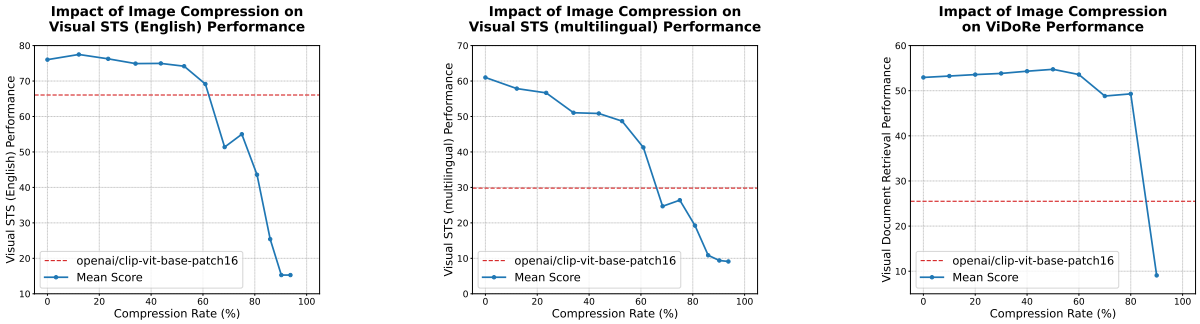

Figure 2: PIXEL LINGUIST II performance under visual token compression.

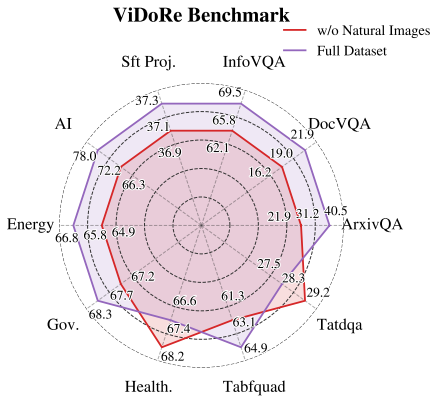

Figure 3: Performance comparison of models trained with and without natural images (Full results in Table 7).

out sacrificing downstream fidelity.

Validation of Multimodal Grounding at Scale Having established in Section 2.2 that purely synthetic pretraining triggers representation collapse, we investigate whether massive data scaling can overcome this limitation. We compare two full-scale variants of PIXEL LINGUIST II pretrained on *Text Corpus 2* with or without natural images (26M LAION pairs).

While this text-only variant matches the full model on fixed-resolution synthetic tasks like Visual STS (Figure 7), its performance drops substantially on the complex ViDoRe benchmark (Figure 3). This full-scale degradation confirms that multimodal grounding cannot be bypassed through scale alone. Natural images act as a foundational regularizer; their diverse layouts, variable aspect ratios, and real-world contexts remain strictly required for robust document understanding at any scale.

6 Related Work

Traditional dual-encoder models like CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023; Tschannen et al., 2025) align images with tokenizer-based text encoders. Previous work demonstrated that language supervision injected certain OCR-related capability into CLIP vision encoders (Tong et al., 2024; Xiao et al., 2025). In parallel, vision-only approaches like PIXEL (Rust et al., 2022) and CLIPPO (Tschannen et al., 2023) model text visually. However, PIXEL relies on reconstruction objectives which lag in semantic discriminability, while CLIPPO lacks native resolution support essential for document processing.

Recent vision-centric retrievers like ColPali (Faysse et al., 2025) utilize late-interaction mechanisms for SOTA performance but incur high latency. Web-SSL (Fan et al., 2025) scales unsupervised visual learning, proving on-par with language supervision. Distinct from these, PIXEL LINGUIST II learns a unified, compact dense vector representation. It achieves high performance on representation benchmarks like ViDoRe while being able to serve as a generalist vision encoder to train MLLMs. Last but not least, PIXEL LINGUIST II proves to serve as a robust and effective vision encoder for emerging trends of context compression (Wei et al., 2025; Cheng et al., 2025).

7 Conclusion

We presented PIXEL LINGUIST II, a unified vision encoder for learning text representations directly from pixels. Rather than treating pixel-text modeling as a matter of scale alone, we identified four design fundamentals: spatial proxies from variable image resolutions and rendered font sizes, multimodal grounding with natural image–text pairs, layout-aware rendering to suppress visual shortcuts, and a multilingual curriculum that separates optical pretraining from semantic alignment. PIXEL LINGUIST II achieves state-of-the-art results on Visual STS and ViDoRe, transfers to downstream MLLM evaluation, and remains robust under aggressive visual token compression.

Limitations

Despite its strong performance, PIXEL LINGUIST II is trained at a smaller data scale than many CLIP- and SigLIP-style baselines that rely on billion-scale image–text corpora such as LAION and DataComp (Schuhmann et al., 2022; Gadre et al., 2023). Scaling the unified pixel-text training recipe, especially its natural image–text component, may further improve visual grounding. In addition, PIXEL LINGUIST II is strongest on dense text and structured documents, but remains less competitive on diagram-heavy scientific subsets such as `ArxivQA`; this suggests that more diverse scientific figures, plots, and diagram-caption pairs would be useful pretraining data. Finally, our rendered text pairs provide controllable layout diversity, but may not cover all noise patterns in real-world text images, such as scans, blur, occlusion, handwriting, and low-quality camera captures.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, and 1 others. 2025. Glyph: Scaling context windows via visual-text compression. *arXiv preprint arXiv:2510.17800*.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, and 1 others. 2023. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, and 1 others. 2025. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and 1 others. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and 1 others. 2021. Openclip. *Zenodo*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document

images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Phillip Rust, Jonas F Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2022. Language modelling with pixels. *arXiv preprint arXiv:2207.06991*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

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

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Michael Tschannen, Basil Mustafa, and Neil Houlsby. 2023. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11017.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. 2025. Mieb: Massive image embedding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22187–22198.

Chenghao Xiao, Zhuoxu Huang, Danlu Chen, G Thomas Hudson, Yizhi Li, Haoran Duan, Chenghua Lin, Jie Fu, Jungong Han, and Noura Al Moubayed. 2024. Pixel sentence representation learning. *arXiv preprint arXiv:2402.08183*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2025. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

A Implementation Details of PIXEL LINGUIST II

We implement DDP with deepspeed ZeRO 2. All-gather is performed across all GPUs and nodes for computing the InfoNCE loss (Oord et al., 2018), which is then sent back to each GPU for back-propagation. We use a global batch size of 32768 facilitated by 64 GPUs and a per-device batch size of 512, and a temperature of 0.03.

B Fonts

We utilize a diverse set of fonts to ensure the robustness of our rendering pipeline. The font library consists of 393 unique font files spanning multiple scripts and weights (100–900). We distribute more kinds of fonts to the language that occupies a larger portion in the dataset. A detailed breakdown of the font families is provided in Table 6.

Category	Font Families
Chinese	Noto Sans/Serif SC, Liu Jian Mao Cao, Long Cang, Ma Shan Zheng, Zhi Mang Xing, ZCOOL Series
Japanese	Noto Sans/Serif JP, DotGothic16, Kiwi Maru, Potta One, Reggae One, RocknRoll One
Korean	Noto Sans/Serif KR, Gothic A1, Do Hyeon, Jua, Yeon Sung, Nanum Series
Arabic	Noto Sans Arabic, Amiri, Cairo
English	Roboto, Open Sans, Lato, Montserrat, Nunito, Playfair Display, Poppins, Quicksand, Raleway, PT Sans, Ubuntu, Lora, Merriweather, Libre Baskerville, Anton, Josefin Sans, Caveat, Dancing Script, Pacifico, Shadows Into Light, Great Vibes, Allura, Cookie, Courgette, Lobster, Parisienne, Sacramento, Satisfy, Tangerine, Yellowtail
Code	JetBrains Mono, Fira Code, Roboto Mono, Source Code Pro, Ubuntu Mono, Inconsolata, Space Mono
Others	Noto Sans variants covering: Armenian, Bengali, Devanagari, Ethiopic, Georgian, Gujarati, Gurmukhi, Hebrew, Kannada, Khmer, Lao, Malayalam, Math, Mongolian, Myanmar, Tamil, Telugu, Thai, and Symbols.

Table 6: Summary of Font Families

C Task Explanation

We display the example of Visual STS and Visual Document Retrieval in Figure 4 and Figure 5.

D Training Configuration

D.1 Render Engine in Pretraining

Table 7 summarizes the quantitative configurations used in our data generation engine.

D.2 Training Setting in End-to-end Evaluation on Downstream Tasks

In our experiments, we utilize various ViT architectures as visual encoders, paired with Qwen2.5-7b-Instruct (Qwen et al., 2025) as the LLM backbone. Following the training paradigm proposed in LLaVA, the training process is divided into two stages:

Stage 1: Only the projector is trainable (the backbones are frozen). We train the model for one epoch with a learning rate of 2.0×10^{-4} and a batch size of 128. A cosine learning rate schedule with a warm-up ratio of 0.1 is applied.

Stage 2: The entire model is fully fine-tuned. We train for three epochs with a learning rate of 2.0×10^{-5} and a batch size of 8, using 4 gradient accumulation steps. Consistent with the first stage, we employ a cosine scheduler with a warm-up ratio of 0.1.

We evaluate the resulting MLLMs on a suite of widely adopted benchmarks covering both text-centric and general multimodal understanding. Specifically, **OCR and document understanding tasks** include InfoVQA (Mathew et al., 2022), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), and

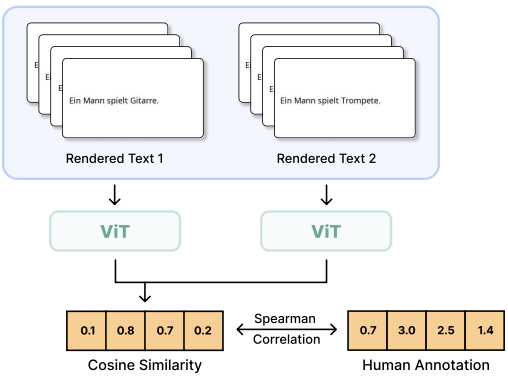

Figure 4: Visual STS task. The text will be rendered as image for models to quantify their semantic equivalence.

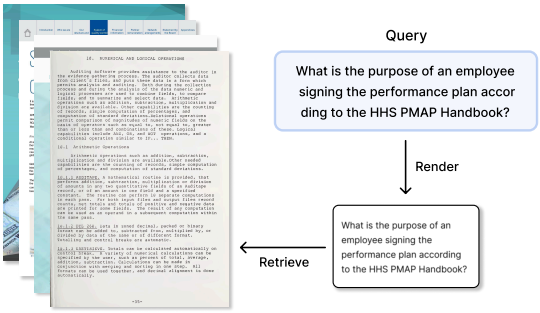

Figure 5: Visual document retrieval (VDR) task. Note that Pixel Linguist processes VDR tasks in an unified way, i.e., the text queries are also first rendered as images and processed in the visual fashion, as opposed to CLIP-style models we benchmark against.

Parameter	Value/Range	Description
Canvas Size	224×224	Fixed input resolution
Font Size	$U(16, 28)$	Sampled uniformly
Max Lines	12	Text wrapping limit
Background Prob. (p_{bg})	0.5	Probability of using DTD textures
Rotation Angle	$[-15°, +15°]$	Random rotation
Position Jitter	± 20 pixels	Random (x, y) shift from center
Blur Prob. (p_{blur})	0.2	Probability of Gaussian Blur
Blur Radius	$[0.5, 1.2]$	Strength of blur
Stroke Prob. (p_{stroke})	0.4	Probability of adding text outline
Brightness Jitter	$[0.6, 1.4]$	Applied to background images

Table 7: Hyperparameters for On-the-fly Text Rendering.

LiveXivVQA (Zhang et al., 2025). **General vision understanding tasks** include AI2D (Kembhavi et al., 2016), MMBench^{EN} (Liu et al., 2024), POPE (Li et al., 2023), RealWorldQA, and MMStar (Chen et al., 2024).

Results are summarized in Table 8, where PIXEL LINGUIST II outperforms Qwen2.5-ViT when used as the vision encoder in end-to-end MLLM training and evaluation, across 9 tasks.

E Examples of Rendered Image

We exhibit multiple rendered textual images from dataset, including multiple languages, in Figure 6.

F Comprehensive Comparison with Baselines

The comparison with full baselines are listed in Table 9, Table 10, Table 11, and Table 12.

G Full Without-Image Ablation

Table 7 displays results of our full scale ablation on natural image incorporation, across Visual STS (English, cross-lingual and multilingual), and ViDoRe.

Figure 6: Examples of Rendered Image

Model	InfoVQA	DocVQA	TextVQA	LiveXIVVQA	AI2D	MMB_{en}	POPE	RealWorldQA	MMStar
PIXEL LINGUIST II	**31.1**	**72.0**	**63.3**	**44.6**	**77.3**	**67.2**	**86.9**	**60.0**	**46.7**
Qwen2.5-ViT	28.7	71.3	63.1	44.4	75.9	64.9	86.6	56.6	45.4

Table 8: Performance comparison on downstream tasks under MLLM evaluation.

Model name	v-STS12	v-STS13	v-STS14	v-STS15	v-STS16	v-STS17	v-STS-b	Avg.
google/siglip-base-patch16-224	63.19	55.40	57.99	73.07	67.79	77.78	54.50	64.25
openai/clip-vit-large-patch14	53.89	66.78	55.98	72.03	70.49	75.26	56.74	64.45
google/siglip-base-patch16-256-multilingual	66.62	54.80	59.00	72.65	68.33	80.53	56.29	65.46
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	57.00	62.25	58.62	74.40	70.57	76.69	58.99	65.50
laion/CLIP-ViT-L-14-laion2B-s32B-b82K	57.52	62.75	59.94	74.55	70.61	75.92	59.43	65.82
openai/clip-vit-base-patch16	63.82	63.26	56.99	73.32	68.91	78.18	57.93	66.06
google/siglip-base-patch16-256	65.01	58.02	60.36	74.25	69.09	78.73	57.65	66.16
google/siglip-base-patch16-384	64.62	59.38	61.17	74.34	70.29	79.27	60.28	67.05
google/siglip-large-patch16-256	63.94	59.44	59.35	75.74	71.83	79.21	62.50	67.43
google/siglip-base-patch16-512	64.97	59.10	61.13	75.08	71.27	80.09	62.21	67.69
google/siglip-so400m-patch14-384	61.90	62.95	60.58	76.17	73.48	78.41	62.63	68.02
laion/CLIP-ViT-B-16-DataComp.XL-s13B-b90K	64.19	63.81	62.34	75.48	69.90	80.04	63.51	68.47
EVA02-CLIP-bigE-14	62.24	62.36	62.17	77.41	73.63	80.96	62.85	68.80
laion/CLIP-ViT-g-14-laion2B-s34B-b88K	61.85	66.43	62.32	76.73	72.67	79.88	64.13	69.14
google/siglip-large-patch16-384	66.30	62.08	61.66	77.11	73.27	79.58	66.59	69.51
laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K	62.36	67.64	64.25	77.36	73.48	80.63	63.38	69.87
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	62.81	68.16	65.50	78.67	74.89	79.97	66.54	70.93
EVA02-CLIP-bigE-14-plus	63.36	68.00	66.38	79.45	75.26	82.87	68.59	71.99
Backbone								
Qwen2.5-VIT	47.50	36.49	30.95	54.69	53.71	63.87	38.63	46.55
Ours								
PIXEL LINGUIST II (mid-training only)	65.78	70.00	67.76	82.39	76.99	84.83	75.30	74.72
PIXEL LINGUIST II (mid-training + finetuned)	76.60	75.94	75.07	85.17	79.65	85.25	80.93	79.80

Table 9: PIXEL LINGUIST II Encoder Performance on Visual STS Tasks (English-only), which renders traditional STS tasks in NLP as image-only tasks, assessing vision models' text-on-image semantic understanding.

Model name	ArxivQA	DocVQA	InfoVQA	Sft Proj.	AI	Energy	Gov.	Health.	Tabfquad	Tatdqa	Avg.
Baselines											
openai/clip-vit-base-patch16	26.54	14.60	51.70	7.13	22.86	32.43	39.84	37.54	17.61	4.71	25.50
google/siglip-base-patch16-224	31.49	16.04	46.11	3.71	25.27	35.53	32.35	37.01	29.04	5.08	26.16
laion/CLIP-ViT-B-16-DataComp.XL-s13B-b90K	28.88	13.97	46.88	7.25	32.17	38.53	31.05	35.83	26.60	9.07	27.02
EVA02-CLIP-bigE-14	32.72	16.35	54.80	10.14	33.53	48.50	41.32	42.98	28.80	7.09	31.62
google/siglip-base-patch16-256	35.17	19.42	48.73	5.45	31.06	41.28	40.07	49.94	37.00	8.50	31.66
EVA02-CLIP-bigE-14-plus	34.86	16.84	55.19	12.76	34.57	44.99	43.14	42.47	30.36	7.52	32.27
openai/clip-vit-large-patch14	28.64	16.69	62.44	17.05	38.25	61.62	52.84	60.23	30.95	11.00	37.97
laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K	34.51	19.68	55.61	16.19	47.20	58.93	50.28	58.04	30.70	15.27	38.64
google/siglip-large-patch16-256	40.19	22.39	54.09	9.13	43.40	50.79	55.45	56.03	49.81	12.38	39.37
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	33.03	19.14	58.82	21.81	54.09	60.23	52.92	55.50	33.11	15.41	40.41
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	38.84	20.44	60.90	25.02	55.42	59.95	62.27	57.86	35.02	16.21	43.19
google/siglip-so400m-patch14-384	50.21	31.28	69.73	25.04	67.78	73.52	75.35	83.10	60.29	27.52	56.38
Ablation											
google/siglip-so400m-patch14-384	50.21	31.28	69.73	25.04	67.78	73.52	75.35	83.10	60.29	27.52	56.38
google/siglip-so400m-patch14-384 (with vision-only paradigm)	20.08	12.05	34.70	8.96	34.54	22.36	26.50	30.09	34.67	11.81	23.58
Δ Performance Difference	30.13 $\downarrow$	19.23$\downarrow$	35.03$\downarrow$	16.08$\downarrow$	33.24$\downarrow$	51.16$\downarrow$	48.85$\downarrow$	53.01$\downarrow$	25.62$\downarrow$	15.71$\downarrow$	32.80$\downarrow$
Backbone											
Qwen2.5-VIT	0.76	0.94	0.74	0.93	0.89	0.00	2.15	1.13	7.15	2.2	1.69
Ours											
PIXEL LINGUIST II (mid-training only)	35.81	20.46	67.61	37.20	75.09	66.30	68.02	67.00	63.96	27.90	52.94
PIXEL LINGUIST II (mid-training + finetuned)	29.87	20.91	69.37	41.60	72.94	73.36	72.77	70.88	71.46	29.36	55.25

Table 10: PIXEL LINGUIST II Encoder performance on Visual Document Retrieval (VDR) Tasks, using ViDoRe benchmark, compared with SOTA baseline encoder models.

Model name	ko-ko	ar-ar	en-ar	en-de	en-tr	es-en	es-es	fr-en	it-en	nl-en	Avg.
openai/clip-vit-base-patch32	18.10	28.30	8.25	22.15	17.97	12.15	47.56	19.48	22.74	25.05	22.18
laion/CLIP-ViT-L-14-laion2B-s32B-b82K	18.23	20.71	4.66	19.38	0.88	19.49	61.89	31.63	27.75	18.38	22.30
laion/CLIP-ViT-B-32-laion2b-s34B-b79K	16.25	21.73	4.20	17.82	17.37	25.07	57.03	22.91	21.49	23.38	22.72
laion/CLIP-ViT-B-16-DataComp.XL-s13B-b90K	19.21	18.40	-1.69	33.07	6.57	16.93	62.39	20.93	19.40	32.23	22.74
EVA02-CLIP-L-14	14.77	29.65	18.89	3.52	16.61	12.23	45.55	32.61	30.84	23.63	22.83
EVA02-CLIP-bigE-14-plus	11.36	31.51	10.71	24.33	-10.05	20.18	59.20	36.12	28.60	33.18	24.52
laion/CLIP-ViT-g-14-laion2B-s34B-b88K	17.17	29.93	14.27	28.50	-4.79	34.19	66.07	29.70	29.02	21.18	26.52
EVA02-CLIP-bigE-14	10.97	29.99	13.49	22.76	6.39	29.03	57.16	36.66	33.43	26.16	26.60
google/siglip-base-patch16-224	21.00	25.03	14.36	31.20	24.80	29.32	69.85	35.70	27.46	28.98	30.77
google/siglip-base-patch16-256	21.40	30.46	12.67	30.19	19.81	28.50	71.68	36.55	28.75	30.72	31.07
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	14.38	32.39	12.21	36.74	14.99	30.44	69.77	39.77	36.44	34.83	32.20
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	19.39	33.39	19.49	43.78	16.68	27.99	62.58	39.32	28.59	37.33	32.85
openai/clip-vit-base-patch16	10.54	36.25	13.13	41.57	35.42	24.63	62.95	38.72	31.40	38.63	33.32
laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K	14.28	36.47	12.75	43.10	19.70	37.37	71.62	36.88	30.78	30.76	33.37
openai/clip-vit-large-patch14	11.07	39.12	18.95	45.71	39.70	36.76	70.11	44.06	40.17	41.63	38.73
google/siglip-so400m-patch14-384	13.65	45.76	11.22	46.07	30.62	40.08	73.62	46.36	36.45	44.95	38.88
Backbone											
Qwen2.5-VIT	51.34	52.45	22.07	24.77	22.58	16.71	65.44	32.05	26.00	26.43	33.98
Ours											
Mid-Training only											
PIXEL LINGUIST II	51.13	50.96	2.09	66.00	54.59	55.51	70.66	63.00	61.91	62.46	53.83
PIXEL LINGUIST II (finetuned)	49.95	43.85	6.96	63.40	50.98	58.19	76.38	62.17	62.22	62.73	53.68
Pretraining + Mid-Training											
PIXEL LINGUIST II	57.51	50.13	1.44	67.4	55.86	61.42	75.82	67.73	67.88	66.36	57.16
PIXEL LINGUIST II (finetuned)	57.24	51.61	3.59	68.55	49.70	64.23	80.48	67.11	64.19	65.39	57.21

Table 11: PIXEL LINGUIST II Encoder Performance on Visual STS Tasks (Cross-lingual Tasks)

Model name	de	es	fr	it	nl	pl	pt	ru	zh	Avg.
openai/clip-vit-base-patch16	32.72	30.81	39.06	29.46	23.46	28.15	26.30	14.69	11.85	26.28
EVA02-CLIP-B-16	30.68	27.02	36.05	27.13	29.71	32.41	29.06	25.40	16.71	28.24
laion/CLIP-ViT-B-32-laion2b-s34B-b79K	41.43	26.40	35.96	28.13	29.75	34.85	28.60	21.84	19.50	29.61
laion/CLIP-ViT-L-14-laion2B-s32B-b82K	39.99	31.22	40.69	28.57	28.49	27.58	25.85	22.66	22.58	29.74
EVA02-CLIP-bigE-14	37.10	35.37	41.49	31.98	28.04	25.33	30.62	25.35	14.58	29.98
laion/CLIP-ViT-B-32-DataComp.XL-s13B-b90K	38.22	28.92	38.00	23.87	32.90	43.21	28.62	27.29	13.95	30.55
openai/clip-vit-large-patch14	37.50	44.18	47.53	36.89	32.51	23.41	35.49	14.06	12.12	31.52
EVA02-CLIP-bigE-14-plus	31.96	37.53	46.88	38.94	29.78	27.50	33.35	25.05	16.20	31.91
laion/CLIP-ViT-B-16-DataComp.XL-s13B-b90K	41.25	31.76	45.92	34.60	35.79	40.38	36.57	26.67	15.18	34.24
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	41.31	39.11	48.44	34.22	34.48	33.20	32.09	26.94	23.88	34.85
google/siglip-base-patch16-224	40.38	41.80	45.75	37.90	37.64	42.65	37.01	32.81	10.79	36.30
laion/CLIP-ViT-g-14-laion2B-s34B-b88K	48.01	41.47	45.03	37.56	36.84	36.02	32.73	30.53	23.65	36.87
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	38.00	43.63	52.36	44.84	34.84	33.19	37.51	28.43	19.19	36.89
google/siglip-base-patch16-256	42.40	44.36	46.72	41.73	38.72	42.34	39.56	35.01	9.34	37.80
laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K	47.05	45.13	50.76	44.24	38.21	34.94	37.87	30.89	14.65	38.19
google/siglip-large-patch16-384	55.72	56.23	54.78	54.24	42.45	41.24	51.62	36.86	14.97	45.35
Backbone										
Qwen2.5-VIT	48.73	45.33	49.35	44.57	40.54	49.01	43.62	48.37	47.16	46.30
Ours										
Mid-Training only										
PIXEL LINGUIST II	64.56	61.74	67.21	63.24	60.14	59.38	58.90	60.78	60.19	61.79
PIXEL LINGUIST II (finetuned)	66.51	66.28	69.10	67.11	63.20	59.82	62.78	61.39	66.79	64.78
Pretraining + Mid-Training										
PIXEL LINGUIST II	66.94	66.03	70.07	66.73	65.12	64.34	62.97	62.18	63.04	65.27
PIXEL LINGUIST II (finetuned)	68.56	70.18	72.21	69.95	66.51	62.84	67.91	64.76	68.27	67.91

Table 12: PIXEL LINGUIST II Encoder Performance on Visual STS Tasks (Multilingual Tasks)

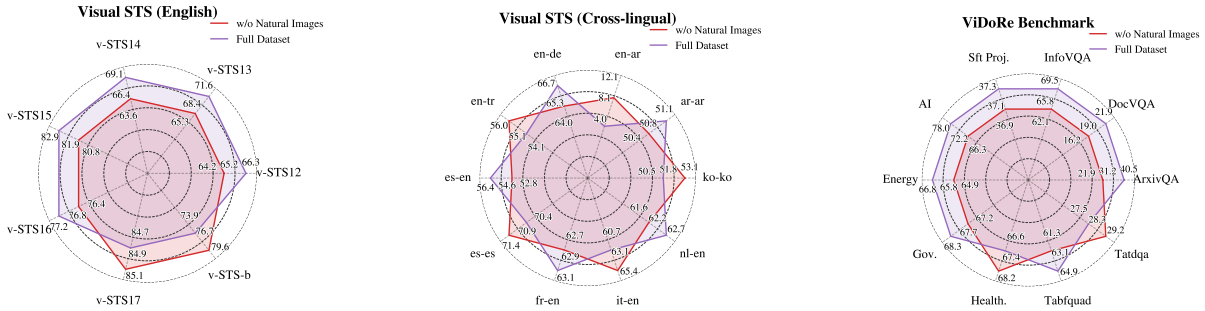

Figure 7: Performance comparison of models trained with and without natural images.