Autoregressive Pre-Training on Pixels and Texts

Yekun Chai[1], Qingyi Liu[2], Jingwu Xiao[3], Shuohuan Wang[1], Yu Sun[1], Hua Wu[1]

[1] Baidu Inc. [2] Sun Yat-sen University [3] Peking University

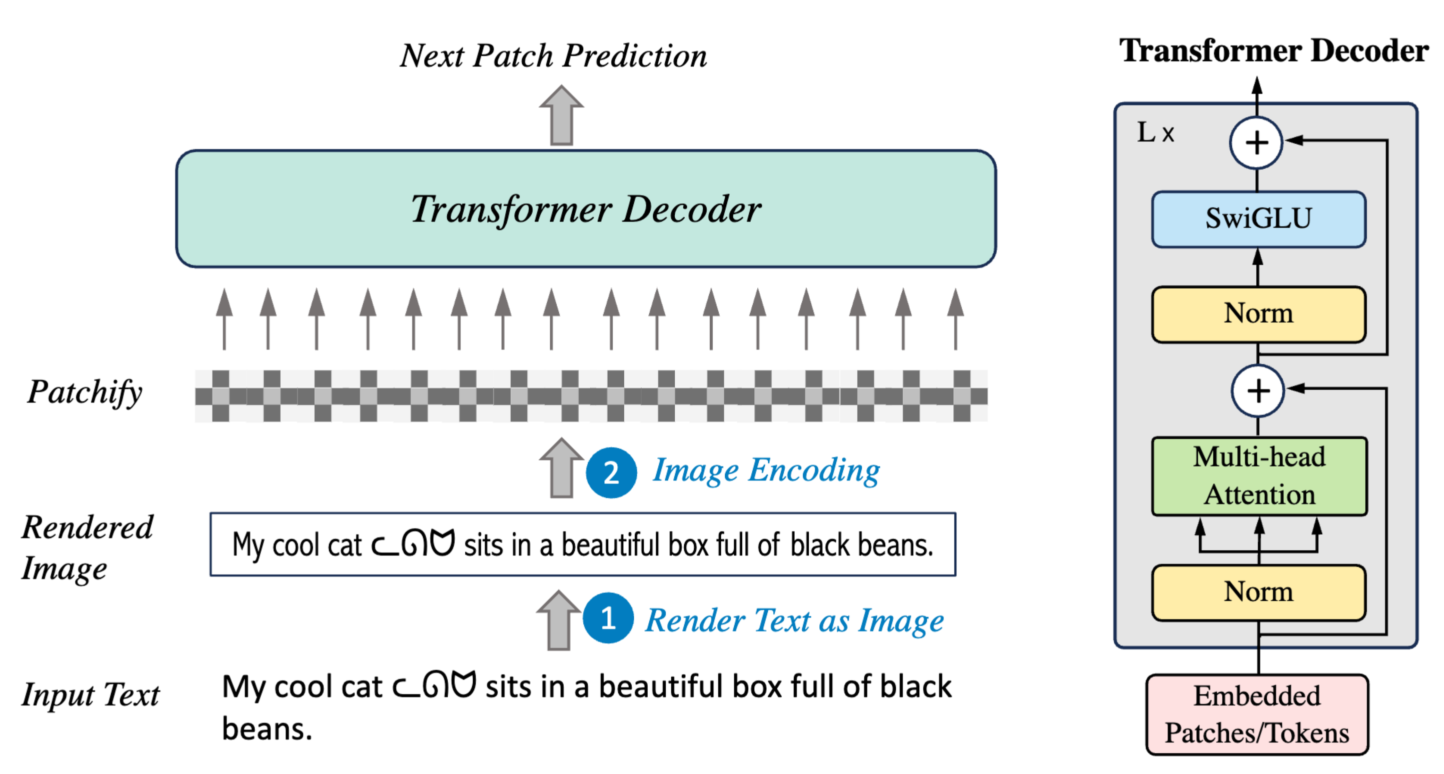
EMNLP 2024

Introduction

The integration of visual and textual information represents a promising direction in the advancement of language models. In this paper, we explore the dual modality of language—both visual and textual—within an autoregressive framework, pre-trained on both document images and texts. Our method employs a multimodal training strategy, utilizing visual data through next patch prediction with a regression head and/or textual data through next token prediction with a classification head.

We focus on understanding the interaction between these two modalities and their combined impact on model performance. Our extensive evaluation across a wide range of benchmarks shows that incorporating both visual and textual data significantly improves the performance of pixel-based language models. This work uncovers the untapped potential of integrating visual and textual modalities for more effective language modeling. We release our code, data, and model checkpoints at https://github.com/ernie-research/pixelgpt.

Visual Text Processing



(a) Visual text image pre-training (*PixelGPT*). (b) Model architecture.

Fig 1. Illustration of pixel-based autoregressive pre-training (PixelGPT).

Pixel Input Preprocessing

① **Text rendering.** Utilize text renderer by converting texts into a visually-rich RGB images.

② **Image encoding**. Split rendered images into patches as in vision transformers.

③ **Autoregressive Training.** Predict next patch based on its historical patches.

Autoregressive Pixel-Text Pretraining

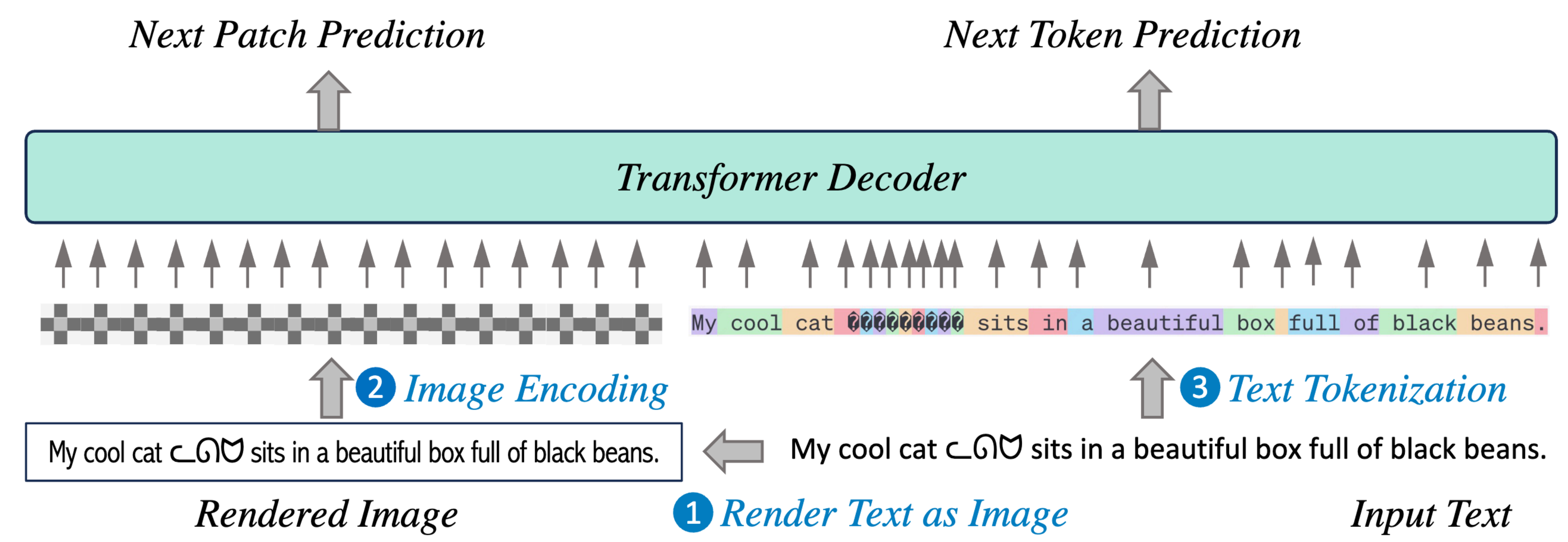


Fig 2. Autoregressive pixel-text pre-training (DualGPT).

Pretraining Objectives

- **Image**: Next patch prediction. Given a sequence of N visual patches $x\_p = (x_p^1, x_p^2, \dots, x_p^N)$ where each visual patch x_t^p is a flattened patch embedding. We decompose the image patch sequence into the production of N conditional probabilities. We use a normalized mean squared error (MSE) loss quantifies the pixel reconstruction accuracy by comparing the normalized target image patches with reconstructed outputs :

$$p(x_p^1, x_p^2, \cdots, x_p^N) = \prod_{t=1}^{N} p(x_p^t | x_p^1, x_p^2, \cdots, x_p^{t-1})$$

- **Text**: Next token prediction. We optimize a cross-entropy loss that evaluates the fidelity of predicted token sequences generated via teacher-forcing against the ground truth tokens.

Pretraining Recipe

- **PixelGPT:** Trained solely on rendered image using MSE loss (Fig 1).
- **MonoGPT:** Trained on separate streams of rendered image and text data without any intermodal pairing.
- **DualGPT**: Trained on unpaired image and text input, and on paired image-text data (dual-modality, Fig 2).

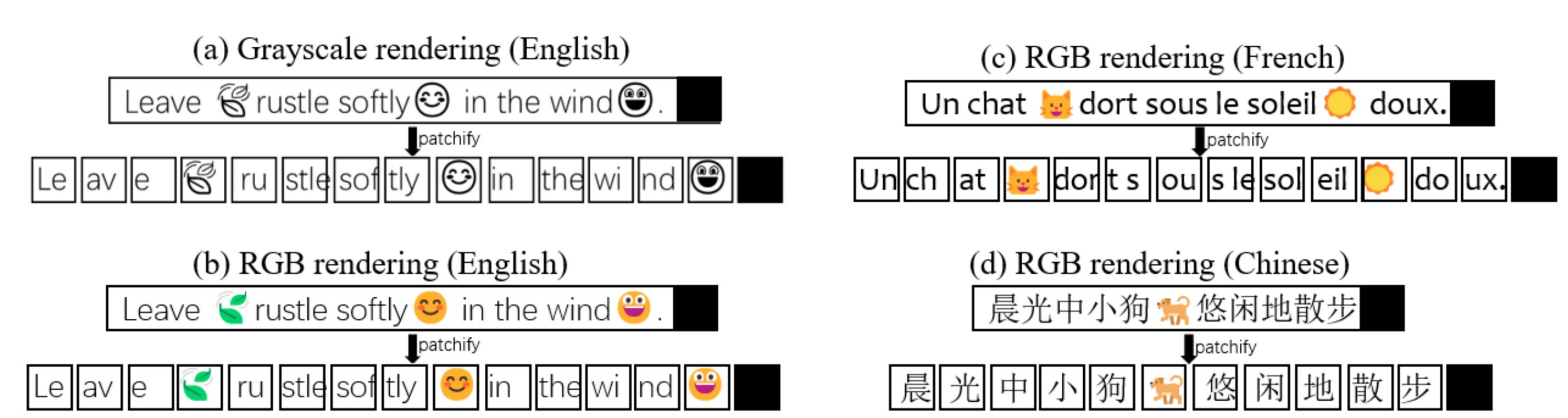


Fig 3. Rendered cases.

Experiments & Analysis

Model	#Param	Input Modality		MNLI-m/mm	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	WNLI	Avg.
		Text	Pixel	Acc	F1	Acc	Acc	MCC	Spear.	F1	Acc	Acc	
BERT	110M	✓	✗	84.0/84.2	87.6	91.0	92.6	60.3	88.8	90.2	69.5	51.8	80.0
GPT-2	126M	✓	✗	81.0	89.4	87.7	92.5	77.0	74.9	71.5	52.0	54.9	75.6
DONUT	143M	✗	✓	64.0	77.8	69.7	82.1	13.9	14.4	81.7	54.9	57.7	57.2
CLIPPO	93M	✗	✓	77.7/77.2	85.3	83.1	**90.9**	28.2	**83.4**	84.5	59.2	-	-
PIXAR	85M	✗	✓	78.4/78.6	85.6	85.7	89.0	**39.9**	81.7	83.3	58.5	**59.2**	74.0
PIXEL	86M	✗	✓	78.1/**78.9**	84.5	**87.8**	89.6	38.4	81.1	**88.2**	60.5	53.8	74.1
PixelGPT	317M	✗	✓	**79.0**/78.2	**86.0**	85.6	90.1	35.3	80.3	84.6	**63.9**	**59.2**	**74.2**

Table 1. Comparative results on GLUE (text vs pixel evaluation).

- **Autoregressive Pixel-based Pre-training Rivals PIXEL**. PixelGPT outperforms PIXEL on QQP (+1.5), RTE (+3.4), and WNLI (+5.4).

Model	#lg	#Param	Input Modality		ENG	ARA	BUL	DEU	ELL	FRA	HIN	RUS	SPA	SWA	THA	TUR	URD	VIE	ZHO	Avg.
			Text	Pixel																
Fine-tune model on all training sets (Translate-train-all)																				
mBERT	104	179M	✓	✗	83.3	73.2	77.9	78.1	75.8	78.5	70.1	76.5	79.7	67.2	67.7	73.3	66.1	77.2	77.7	74.8
XLM-R base	100	270M	✓	✗	85.4	77.3	81.3	80.3	80.4	81.4	76.1	79.7	82.2	73.1	77.9	78.6	73.0	79.7	80.2	79.1
BERT	1	110M	✓	✗	83.7	64.8	69.1	70.4	67.7	72.4	59.2	66.4	72.4	62.2	35.7	66.3	54.5	67.6	46.2	63.9
PIXEL	1	86M	✗	✓	77.2	**58.9**	66.5	68.0	64.9	69.4	57.8	63.4	70.3	60.8	**50.2**	64.0	54.1	64.8	**52.0**	62.8
PixelGPT	1	317M	✗	✓	**77.7**	55.4	**66.7**	**69.0**	**67.4**	**71.2**	**59.1**	**65.6**	**71.4**	**61.7**	47.0	**65.2**	**54.4**	**66.1**	50.5	**63.2**

Table 2. Cross-lingual evaluation on XNLI (*Translate-Train-All*).

- **PixelGPT matches the performance of BERT,** and consistently surpasses the in average accuracy across multilingual XNLI dataset.

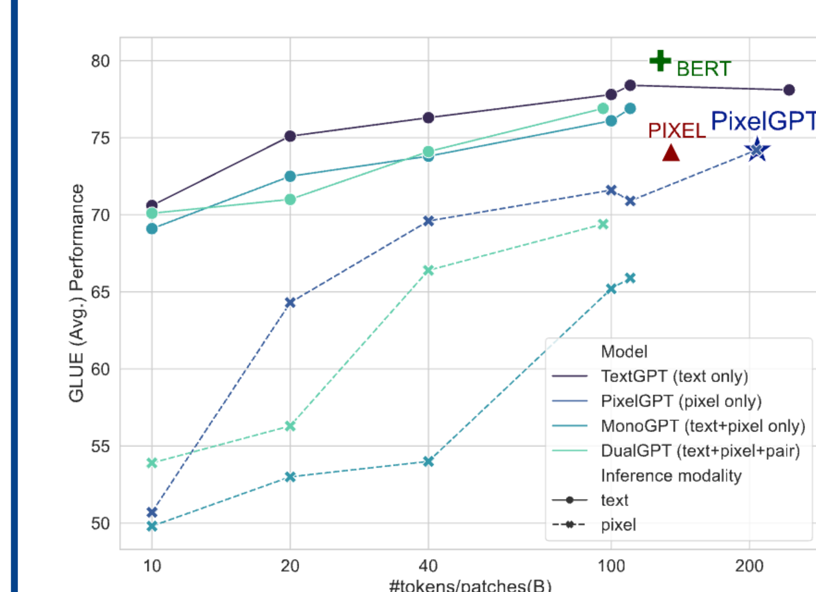


Fig 4. Scaling trend on GLUE.

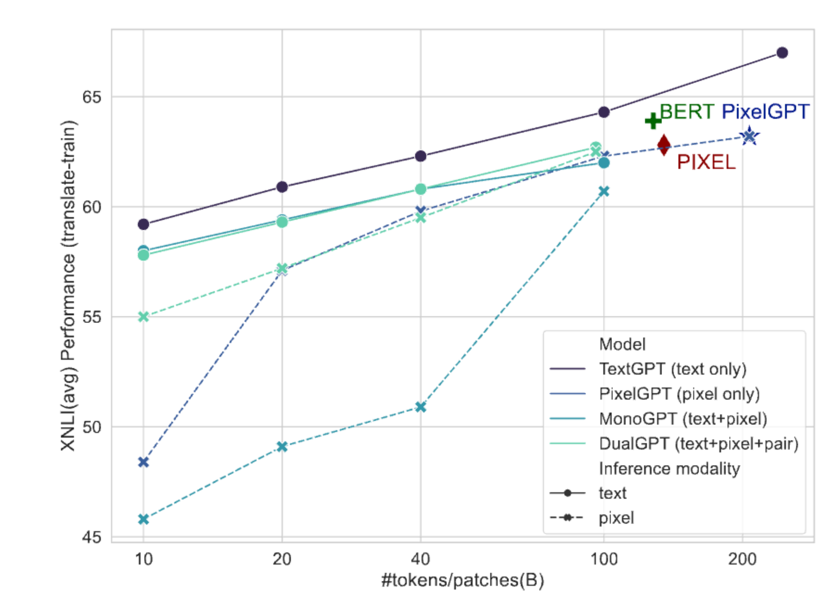


Fig 5. Scaling trend on XNLI.

① Pixel-based training exhibit an increased data demand.

② Utilizing paired dual-modality data improves multimodal learning, particularly for pixel-based input.

Model	Input Modality		MNLI-m/mm	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	WNLI	Avg.
	Text	Pixel	Acc	F1	Acc	Acc	MCC	Spear.	F1	Acc	Acc	
TextGPT (text only)	✓	✗	79.9/80.0	86.1	86.1	91.5	47.3	**85.8**	86.3	63.5	56.3	76.3
MonoGPT (text+pixel)	✓	✗	80.0/**80.5**	85.9	**87.3**	90.1	40.2	83.8	87.0	62.8	56.3	75.4
	✗	✓	64.7/65.9	78.9	77.3	74.8	11.6	73.2	83.5	59.9	57.7	64.8
DualGPT (text+pixel+pair)	✓	✗	**80.1**/80.4	**86.5**	86.8	**91.6**	**49.0**	85.4	**87.6**	65.7	56.3	**76.9**
	✗	✓	71.5/71.7	82.8	81.6	83.4	17.2	80.2	84.1	**66.4**	**59.2**	69.4

Table 3. Ablation study on GLUE.

- Paired dual-modality data improves the language understanding.