m3P: Towards Multimodal Multilingual

Jian ... Wang[1],
Jiaheng L ... Zhoujun Li[1]

m3P: Towards Multimodal Multilingual
Translation with Multimodal Prompt

Jian Yang[1], Hongcheng Guo[1], Yuwei Yin[2], Jiaqi Bai[1], Bing Wang[1],
Jiaheng Liu[1], Xinnian Liang[1], Linzheng Chai[1] , Liqun Yang[1] and Zhoujun Li[1]

[1]State Key Lab of Software Development Environment,
Beihang University, Beijing, China

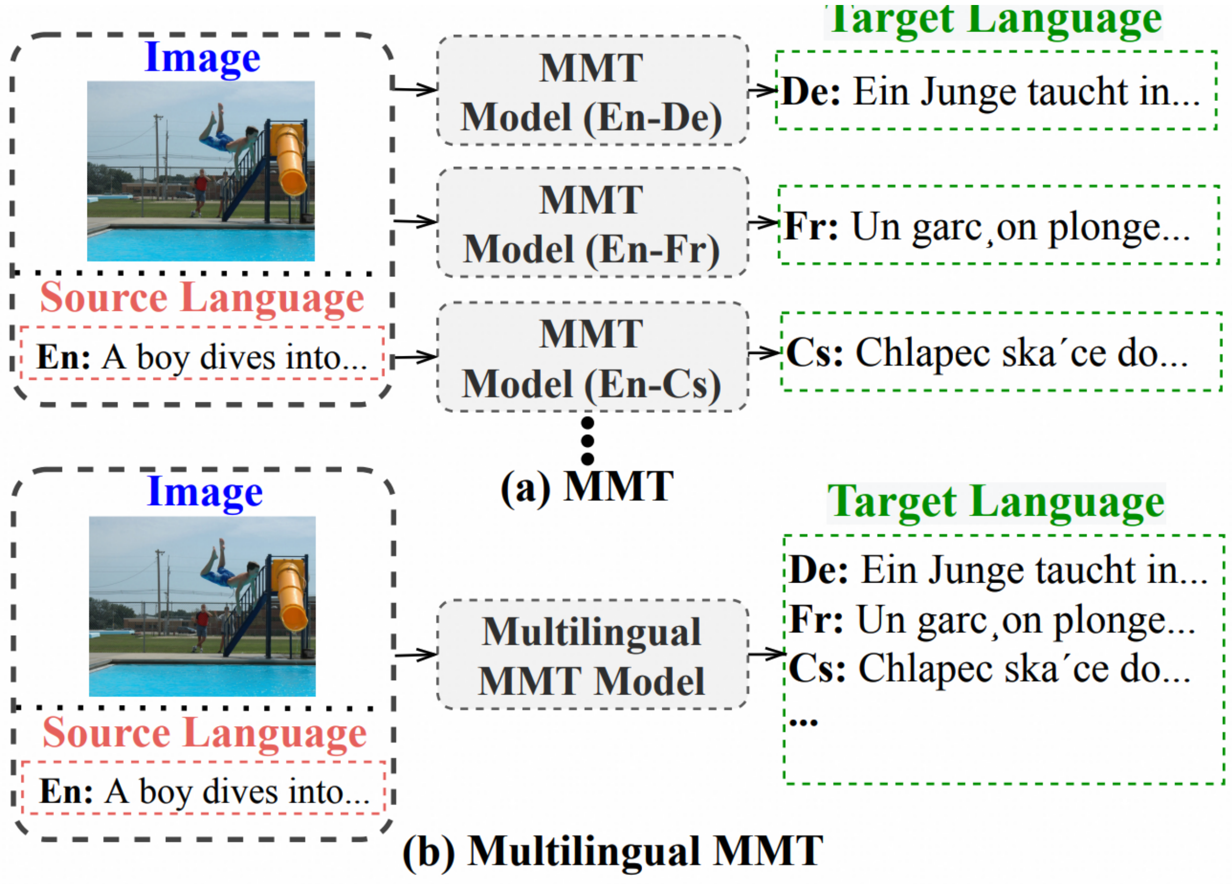
2 Department of Computer Science, University of British Columbia

COLING 2024

Outline

Encoder-Decoder Pre-training

- Multilingual Multimodal Machine Translation
  - Multilingual Machine Translation.
  - Multimodal Machine Translation.

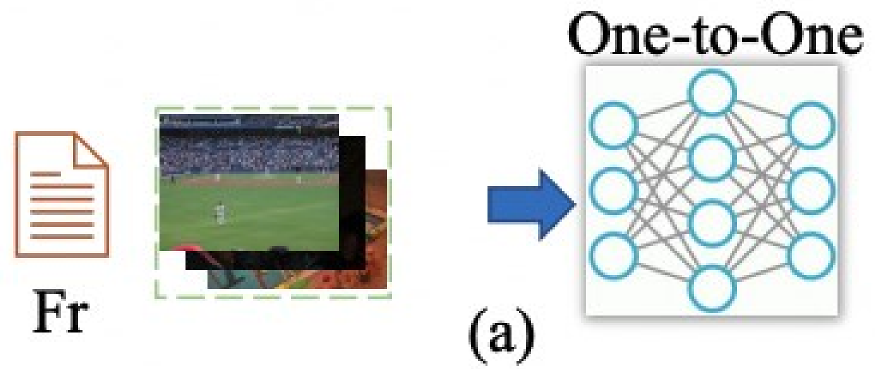


(a) MMT

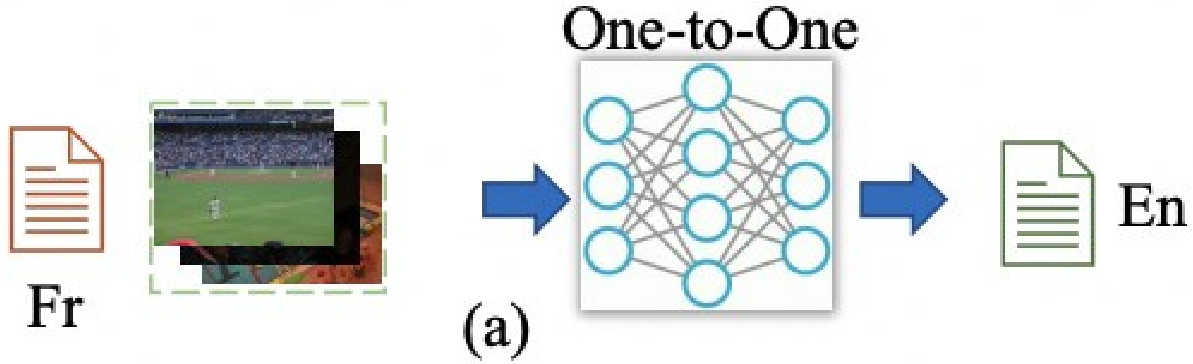
(b) Multilingual MMT

Encoder-Decoder Pre-training

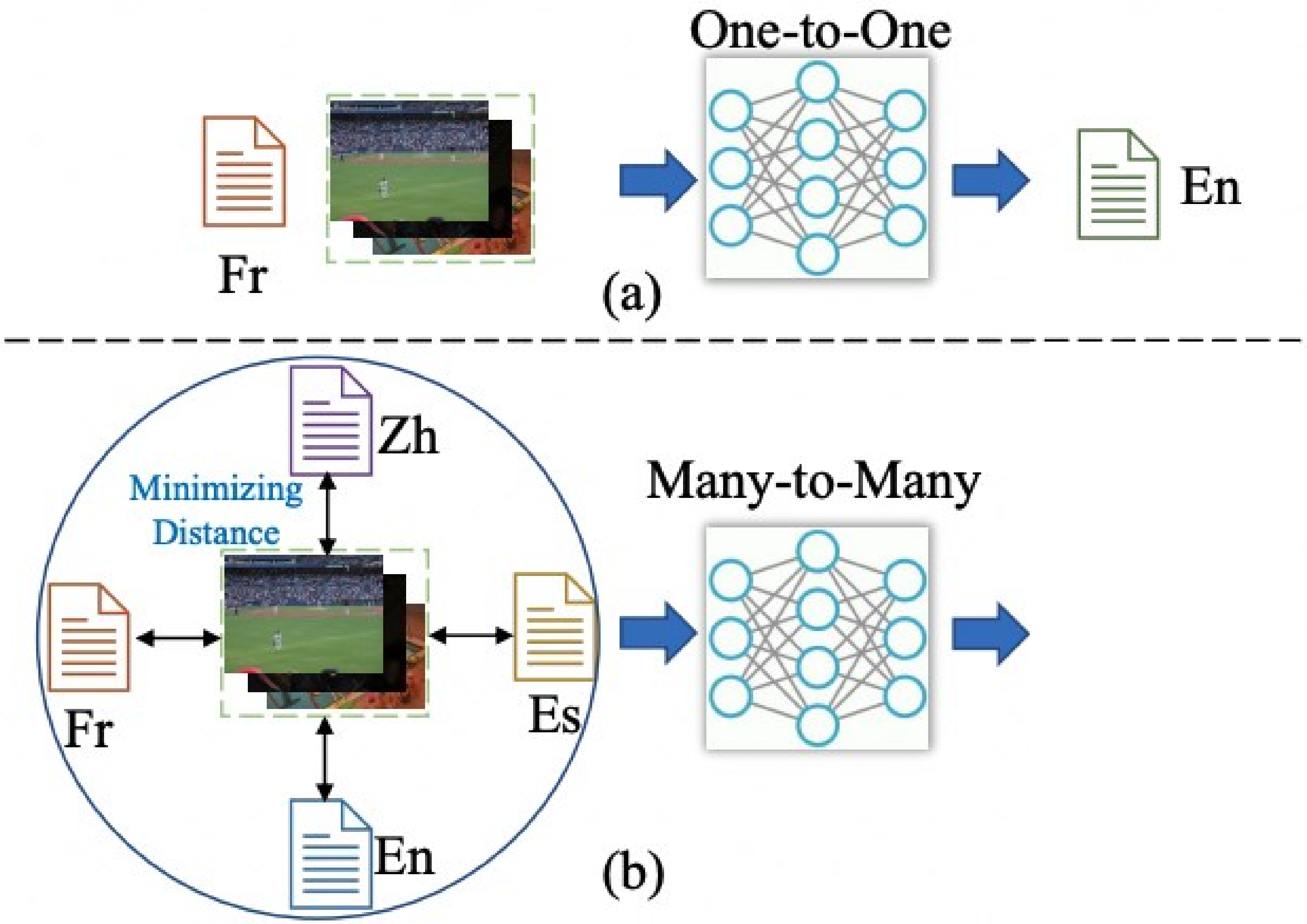
Encoder-Decoder Pre-training



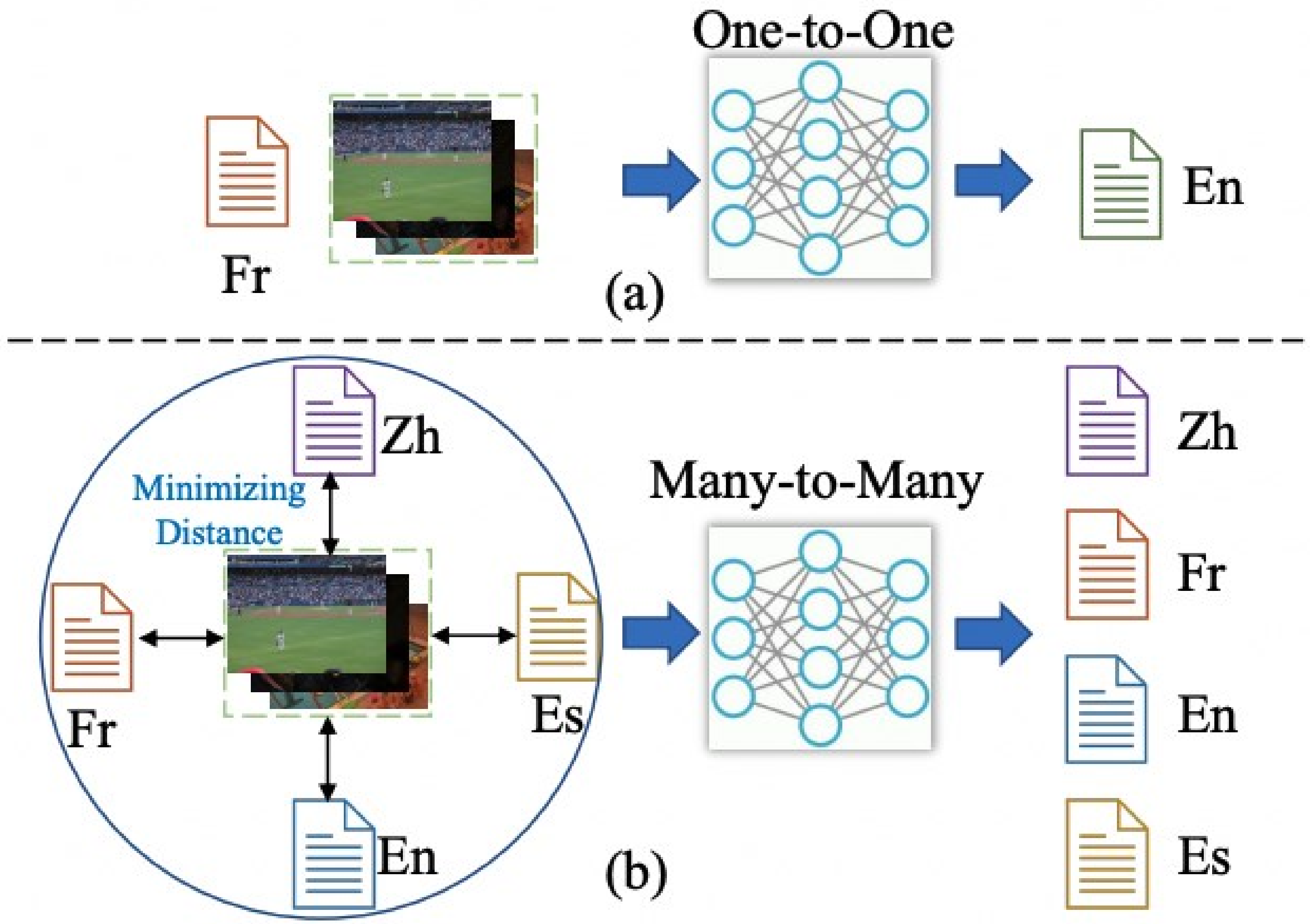
Encoder-Decoder Pre-training



Encoder-Decoder Pre-training



Encoder-Decoder Pre-training



Outline

Multilingual Multimodal Translation

- *Multilingual Multimodal Translation*
  - Given M bilingual corpora with images $D_{all} = \{D_m\}_{m=1}^{M}$, where M denote the number of the training corpora of N languages $L_{all} = \{L_n\}_{n=1}^{N}$ and L_n denote the n-th language. Each bilingual corpus with images $D_m = \{x^k, y^k, z^k\}_{k=1}^{K}$ from D_{all} consists of the source sentences, target sentences, and corresponding images. The training objective of multilingual multimodal translation can be described as:

$$\mathcal{L}_m = -\sum_{m=1}^{M} \mathbb{E}_{x^k, y^k, z^k \in D_m} \left[\log P(y^k | x^k, z^k; \Theta)\right]$$

Multilingual Multimodal Translation

(a) *Decoder-only Prompt:*
{Decoder}
Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction:
Please translate the following sentence from
$\{L_i\}$ to $\{L_j\}$: $\{z^k\}$ $\{x^k\}$
Response:
$\{y^k\}$

(b) *Encoder-Decoder Prompt:*
{Text Encoder}:
Please translate the following sentence from
$\{L_i\}$ to $\{L_j\}$: $\{x^k\}$
{Vision Encoder}:
$\{z^k\}$
{Vision Decoder}:
$\{y^k\}$

Multilingual Multimodal Alignment

- *Multilingual Multimodal Alignment*

$$\mathcal{L}_c = \sum_{x^k, z^k \in D_{all}} \left(f(x^k, z^k) + f(z^k, x^k) \right)$$

$$f(x^k, z^k) = -\log \frac{\exp\left(z^k \cdot x^k / \tau\right)}{\sum_{x \in \{x^k, x^-\}} \exp\left(z^k \cdot z / \tau\right)}$$

$$f(z^k, x^k) = -\log \frac{\exp\left(z^k \cdot x^k / \tau\right)}{\sum_{x \in \{x^k, x^-\}} \exp\left(z^k \cdot z / \tau\right)}$$

Multilingual Multimodal Augmentation

Augmentation

- For image augmentation, we leverage the function $I(\cdot)$ to augment the original image by cropping, resizing, rotation, cutout, color distortion, Gaussian blur, and Sobel filtering. Then, we divide an image into regular non-overlapping patches and mask the chosen patches sampling from a uniform distribution as masked image modeling.
- For the multilingual text, we randomly mask some random spans of contiguous tokens. For each sentence, we adopt the multilingual data augmentation $\mathrm{T}(\cdot)$ to augment the original sentence of different languages. The augmented source sentence and the image $\{\mathrm{T}(x^k),\mathrm{T}(z^k)\}$ with multilingual multimodal augmentation (MMA) is used to enhance the contrastive learning to learn the specific representational invariances.

Multilingual Generation

☐ *Multi-task Training*

$$e^k = \Big\Vert_{a=1}^{A} \sigma\left(\frac{(W_Q^a h^k)(W_Q^a s^k)^\top}{\sqrt{C}}\right)(W_V^a s^k)$$

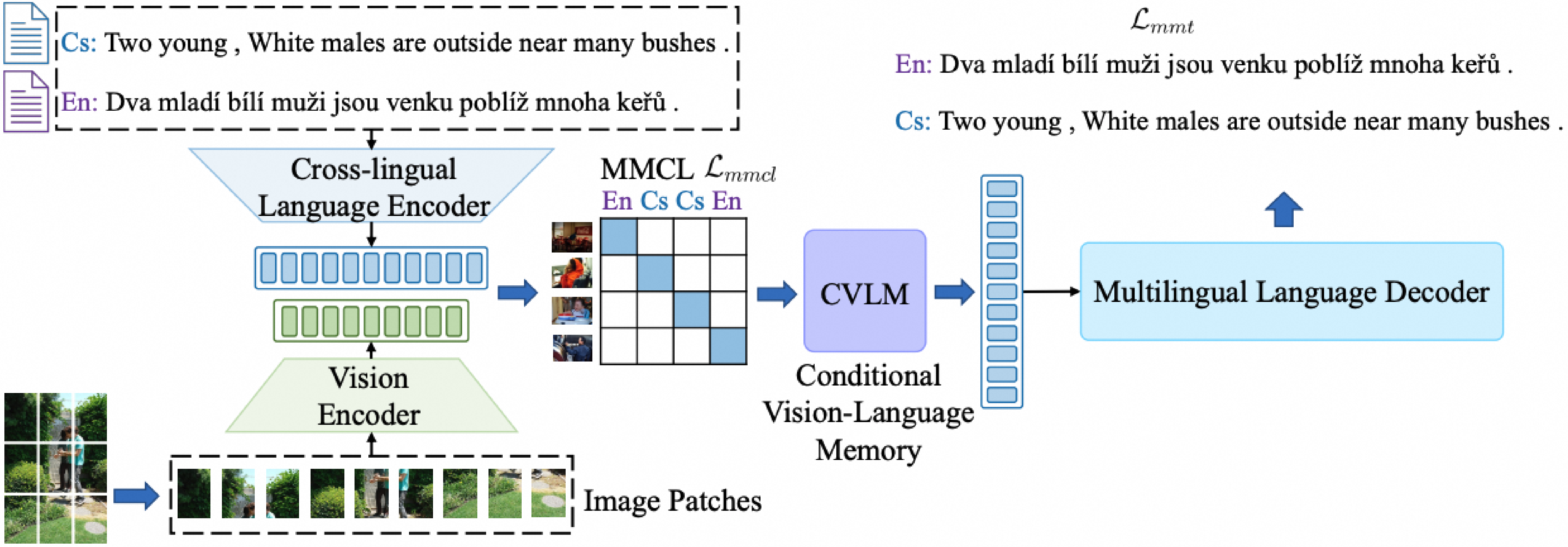
$$y_t^k = \mathcal{D}(y_{1:t-1}^k, s^k; \theta)$$

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, h^k; \theta)$$

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, e^k; \theta)$$

$$\mathcal{L}_{all} = \mathcal{L}_m + \lambda \mathcal{L}_c$$

Model Overview



Outlines

Flickr Test Set

		En→Fr	En→Cs	En→De	Fr→En	Cs→En	De→En	Avg_6
Only Trained on Text Data								
1→1	BiNMT (Vaswani et al., 2017)	63.3	33.4	39.9	54.0	41.1	43.8	45.9
N→N	MNMT (Fan et al., 2021)	63.8	34.0	40.2	52.0	41.3	42.5	45.6
Trained on Text and Vision Data								
1→1	BiNMT (Vaswani et al., 2017)	63.5	33.0	40.3	55.1	41.8	44.1	46.3
N→N	MNMT (Gated Fusion) (Li et al., 2021a)	63.8	34.4	41.0	51.5	41.1	43.3	45.8
	MNMT (Concatenation) (Li et al., 2021a)	63.0	33.8	38.8	53.3	43.6	44.0	46.1
	mRASP2 (Pan et al., 2021)	63.8	34.4	41.3	53.2	44.0	44.5	46.9
	Selective Attn (Li et al., 2022)	63.5	34.4	41.3	53.2	44.0	44.5	46.8
	LVP-M^3 (Guo et al., 2022b)	63.4	34.1	41.4	53.2	44.0	44.5	46.8
	M^3P (Encoder-Decoder)	**64.8**	**35.2**	**41.8**	**53.8**	**44.8**	**45.0**	**47.6**
	M^3P (Decoder-only)	**66.4**	**38.1**	**43.5**	**56.7**	**46.9**	**48.1**	**49.9**

MSCOCO Test Set

- Evaluation on Multilingual Translation and Extractive summarization.

		En→Fr	En→De	De→En	Fr→En	Avg_4	En→Fr	En→De	Fr→En	De→En	Avg_4
		Flick2017					MSCOCO				
Only Trained on Text Data											
1→1	BiNMT (Vaswani et al., 2017)	55.4	34.1	39.2	43.4	43.0	45.8	32.1	40.6	34.3	38.2
N→N	MNMT (Fan et al., 2021)	56.8	34.9	40.3	44.6	44.2	45.9	31.9	41.6	34.6	38.5
Trained on Text and Vision Data											
1→1	BiNMT (Vaswani et al., 2017)	55.8	34.6	39.6	43.6	43.4	45.8	32.3	41.6	34.4	38.5
N→N	MNMT (Gated Fusion) (Li et al., 2021a)	56.8	34.3	40.3	44.2	43.9	46.8	32.5	42.2	34.5	39.0
	MNMT (Concatenation) (Li et al., 2021a)	56.4	34.0	39.4	43.8	43.4	46.4	32.6	42.4	34.1	38.9
	mRASP2 (Pan et al., 2021)	57.0	35.1	39.6	44.1	43.9	47.1	32.7	42.3	34.8	39.2
	Selective Attn (Li et al., 2022)	56.6	34.2	40.3	44.4	43.9	46.8	32.5	42.5	34.3	39.0
	LVP-M^3 (Guo et al., 2022b)	57.4	34.4	40.4	44.7	44.2	46.8	32.5	42.6	34.5	39.1
	M^{3}P (Encoder-Decoder)	**57.4**	**35.3**	**41.0**	**45.6**	**44.8**	**46.8**	**33.1**	**43.2**	**35.2**	**39.6**
	M^{3}P (Decoder-only)	**58.3**	**37.2**	**42.2**	**46.5**	**46.1**	**47.4**	**34.2**	**44.5**	**36.2**	**40.6**

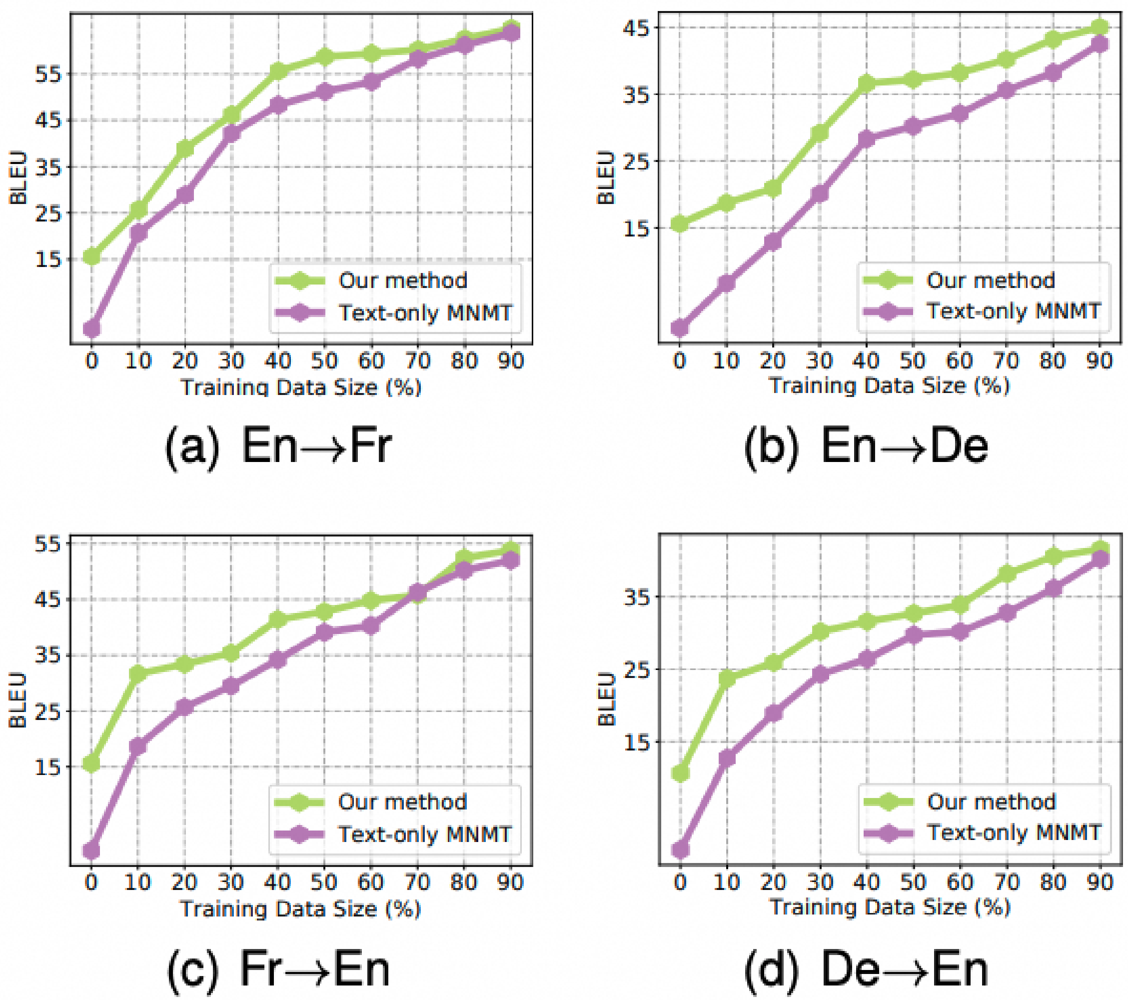
Ablation study

- Evaluation on Multilingual Translation and Extractive summarization.

ID	Flickr2016	En→De	De→En
①	M^{3}P (our method)	**41.6**	**45.0**
②	① - MMCL	41.2	44.6
③	② - CVLM	40.8	44.0
④	③ - MDropNet	40.5	43.8
⑤	④ - Multilingual Training	40.1	43.2

Outlines

Low-resource setting



Ablation Study

(a) Original (b) En (c) De (d) Fr (e) Cs

Outlines

Conclusion and Future Work

Conclusion

- we introduce m 3P, a state-of-the-art multilingual multimodal machine translation model, which supports multiple translation directions of 102 languages guided by image context.
- To narrow the gap among different languages, the image is operated as the central language by contrastive learning (MMCL) trained on the multilingual text-image pairs. Then, we incorporate the visual context into the language representations as the conditional vision-language memory (CVLM) for multilingual generation.
- Extensive experiments prove the effectiveness of m3P on the Multi30k and the extended large-scale dataset InstrMulti102 of 102 languages.