

# Mitigating Reporting Bias in Visual-Language Datasets w/ Large Generative Models

---

QIYU WU

PH.D. STUDENT AT TSURUOKA LAB

THE UNIVERSITY OF TOKYO

Great thanks to Chu Sensei  
for the invitation!

# Contents

---

- About me
- Brief introduction to projects I have done in the past
- **Recent paper introduction**
  - What is reporting bias?
  - How reporting bias can affect the visual-language model.
  - Mitigate reporting bias by decoupling object-attribute association.
- Q&A

# About Me

---

- Qiyu (チーユー) Wu (ウー)
- 3<sup>rd</sup> year PhD Student at The University of Tokyo, advised by Prof. Yoshimasa Tsuruoka
- Visiting student at UCSD, Master from Peking University, Bachelor from Sichuan University
- Former intern at Creative AI Lab @ Sony, MSRA and STCA @ Microsoft and Baidu Research.
- Research interest
  - **Semantic representation learning for language (and beyond)**
  - Casual: data mining and any AI-related things



# Things I have done in the past

---

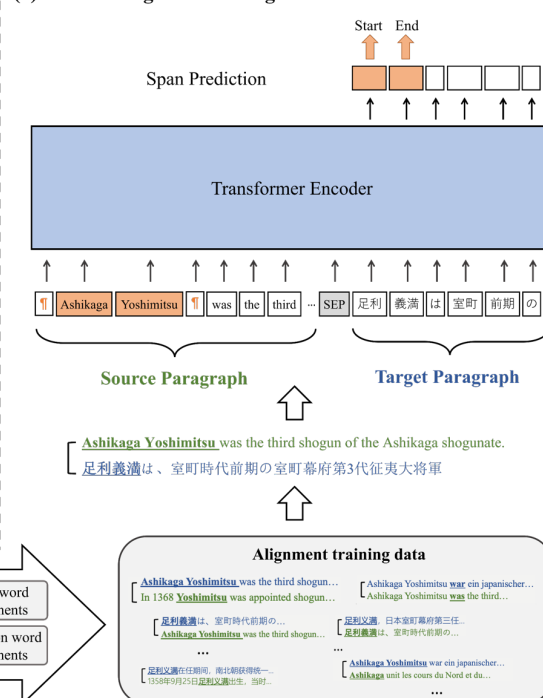
1. WORD/PHRASE LEVEL REPRESENTATION
2. SENTENCE LEVEL REPRESENTATION
3. REPRESENTATION FOR OTHERS

# Things I have done in the past: word/phrase-level representation

## (1) Data Collection and Annotation



## (2) Pre-training for word alignment



## Without Notes:

COVID-19 has cost thousands of \_\_\_\_\_.

What is COVID-19?



dollars?  
donuts?  
puppies?  
tomatoes?

## With Notes:

COVID-19 has cost thousands of lives.



Pandemic;  
global crisis

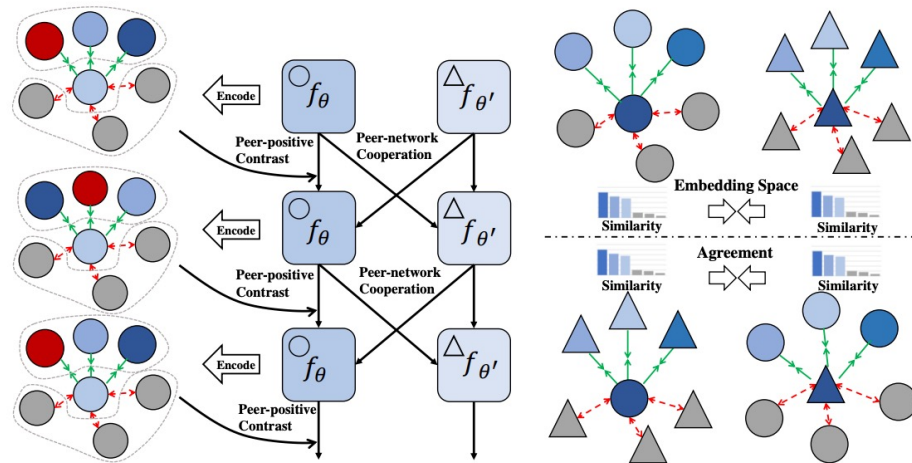
A note of 'COVID-19' taken from a previously seen sentence:  
*The **COVID-19** pandemic is an ongoing global crisis.*

Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training ([Wu et al. ACL 2023](#))

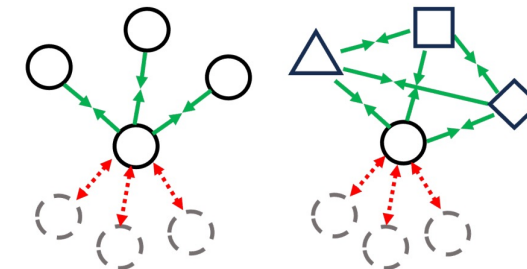
Utilize cross-sentence signal to address rare words issue in language model pre-training ([Wu et al. ICLR 2021](#))

# Things I have done in the past: sentence representation

Augmenting	Order	N-gram	Bag-of-words
<i>Shuffled Sentence</i>	×	×	✓
<i>Inversed Sentence</i>	×	✓	✓
<i>Word Repetition</i>	✓	×	✓
<i>Word Deletion</i>	✓	×	×

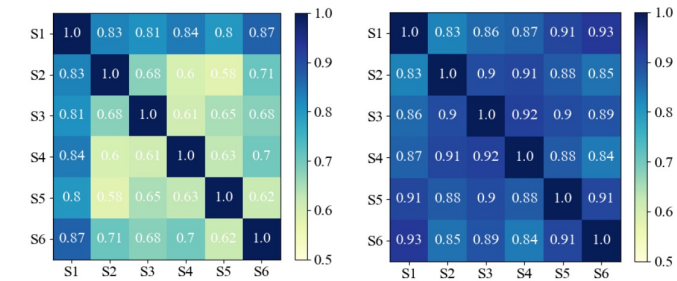


PCL: Unsupervised data augmentation for sentence embedding by contrastive learning ([Wu et al. EMNLP 2022](#))



(a) Mono-lingual

(b) Multi-lingual

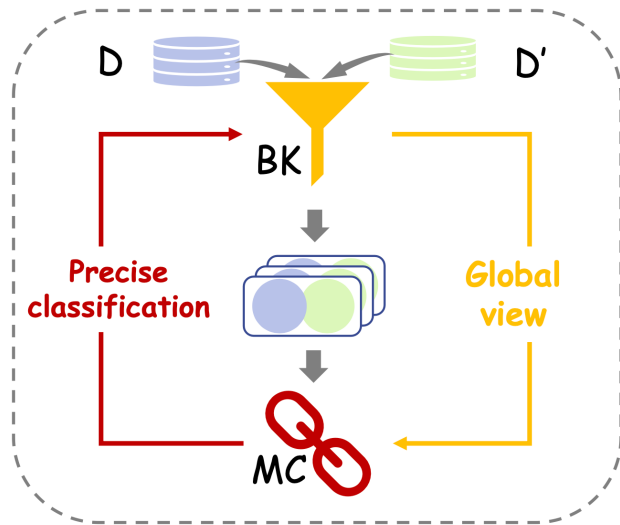


(c) Mono-lingual

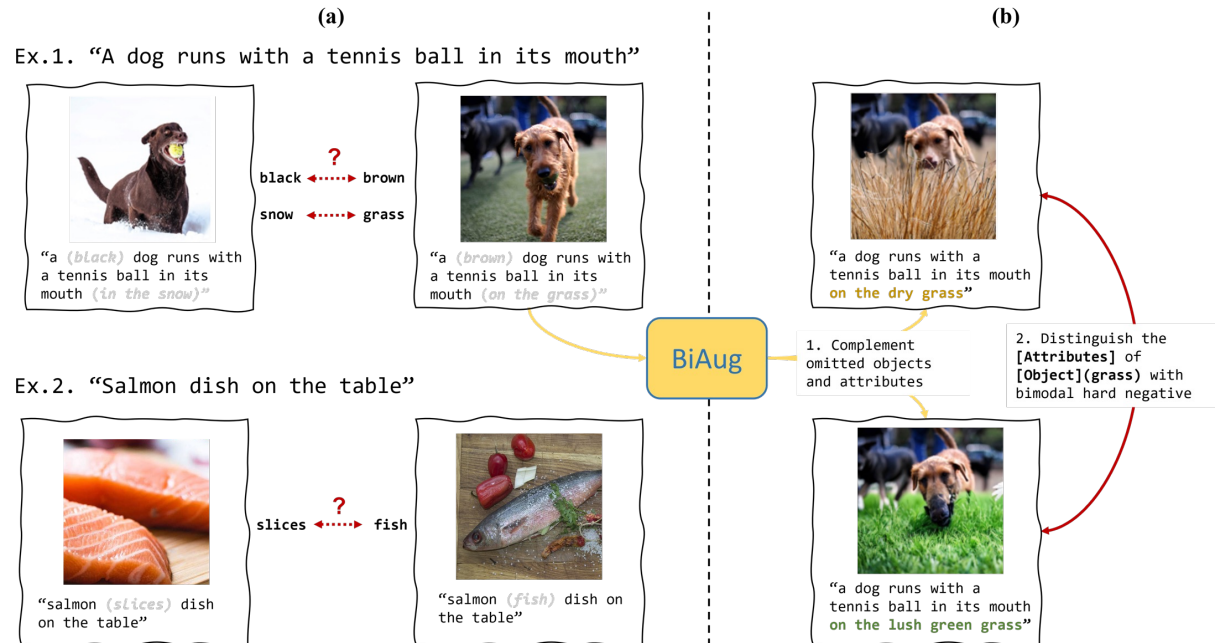
(d) Multi-lingual

MPCL: Diverse positives for multilingual sentence embeddings ([Zhao et al. arXiv 2023](#))

# Things I have done in the past: semantic representation in database and **visual-language** (today's topic!)



CLER: matcher and blocker (entity representation) for entity resolution can benefit mutually ([Wu and Wu et al. to appear at VLDB 2024](#))



**BiAug: bimodal augmentation to mitigate reporting bias in visual-language datasets** ([Wu et al. arXiv 2023](#))

# Question?

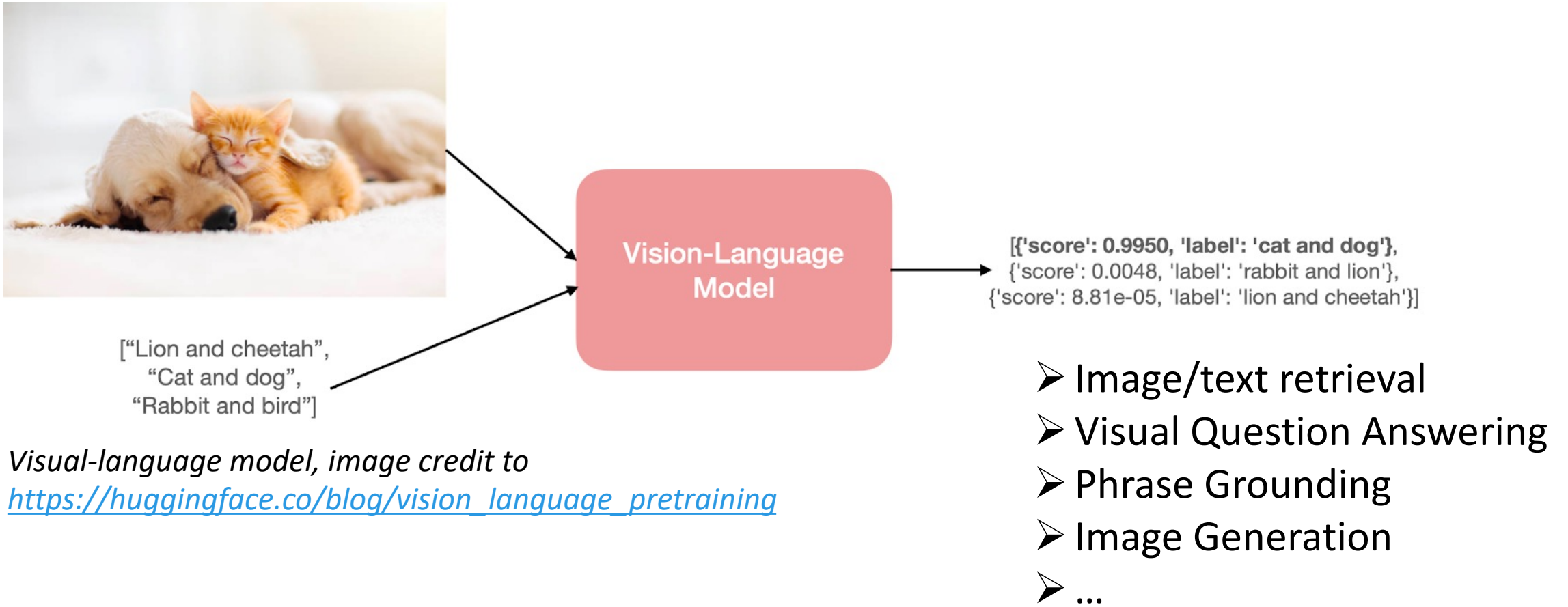
---

# Towards reporting bias in visual-language datasets: bimodal augmentation by decoupling object-attribute association

---

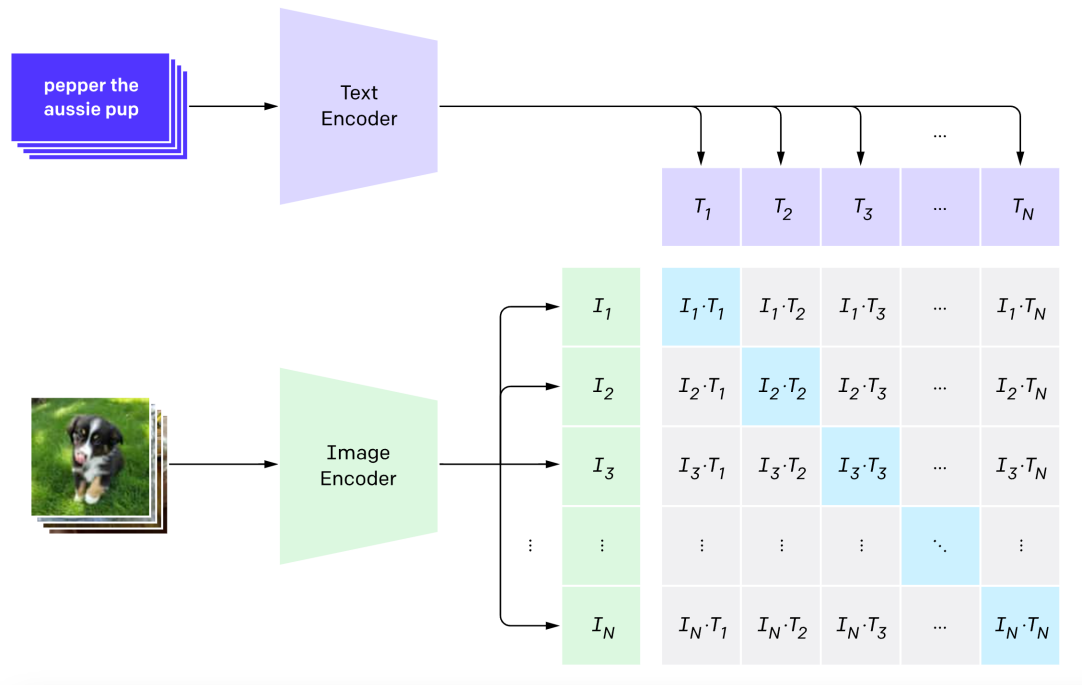
[HTTPS://ARXIV.ORG/PDF/2310.01330.PDF](https://arxiv.org/pdf/2310.01330.pdf)

# Visual-language model is great

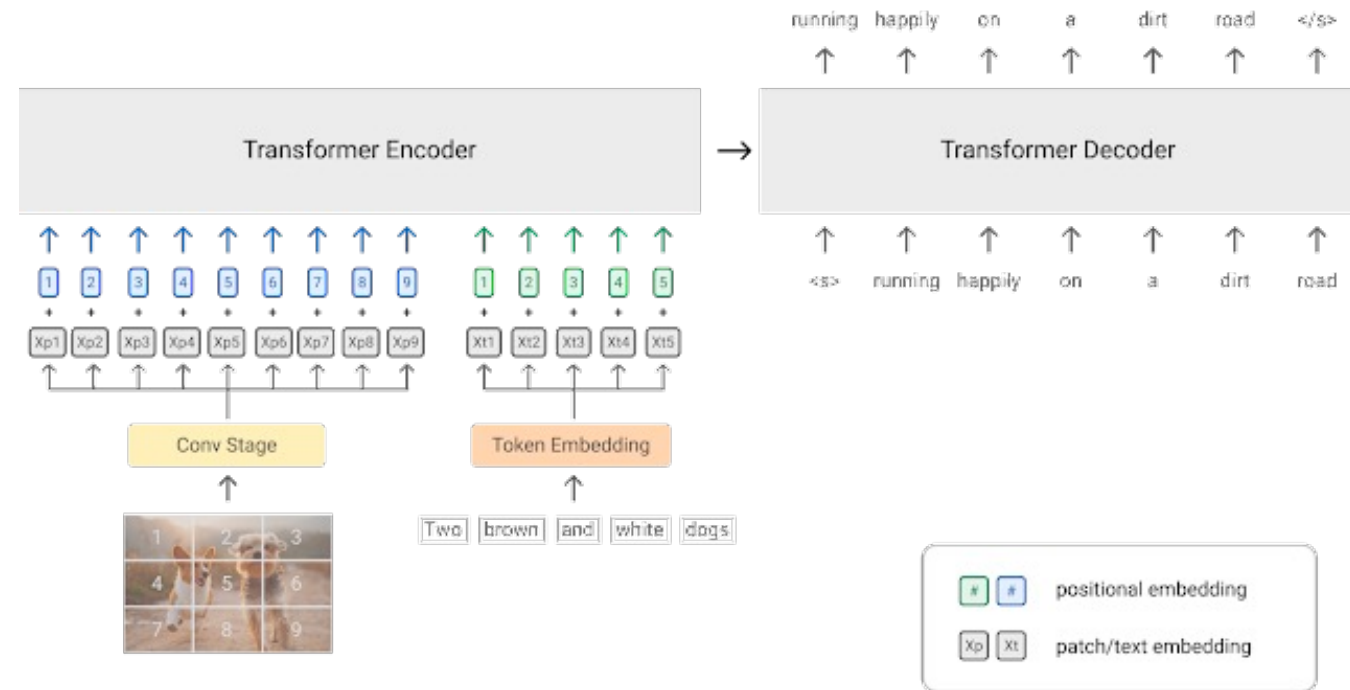


Visual-language model, image credit to  
[https://huggingface.co/blog/vision\\_language\\_pretraining](https://huggingface.co/blog/vision_language_pretraining)

# Learning a visual-language model from image-caption pair



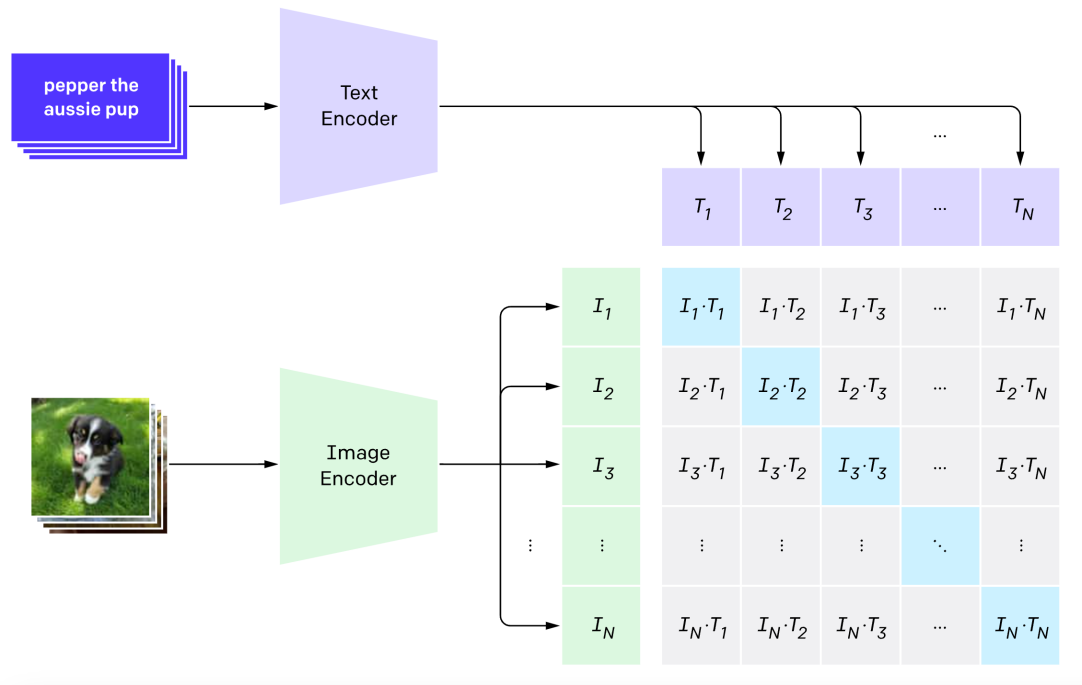
*Contrastive Learning, image credit to CLIP*  
<https://openai.com/research/clip>



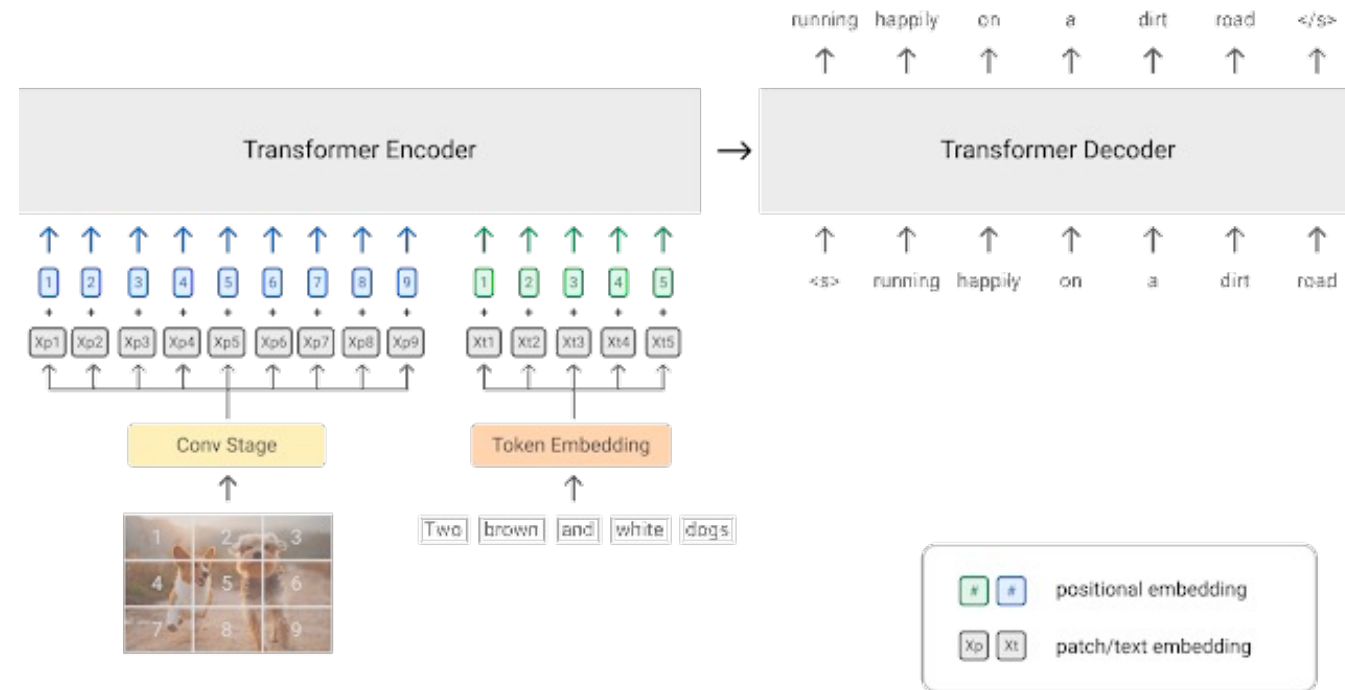
*PrefixLM, image credit to SimVLM*  
<https://arxiv.org/pdf/2108.10904.pdf>



# Learning a visual-language model from image-caption pair which is crucial



*Contrastive Learning, image credit to CLIP*  
<https://openai.com/research/clip>

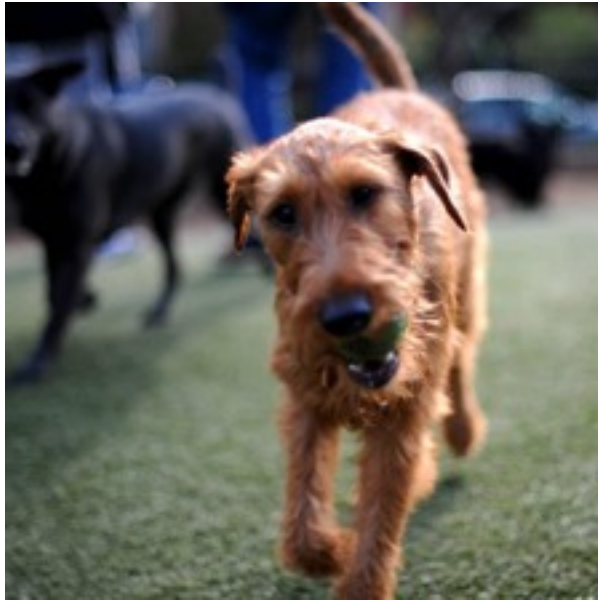


*PrefixLM, image credit to SimVLM*  
<https://arxiv.org/pdf/2108.10904.pdf>

# Visual-language models are learned from image-caption pair like this

---

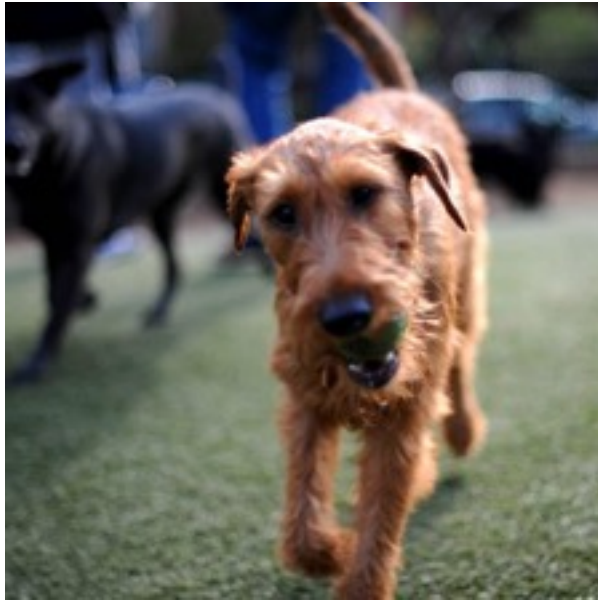
“A dog runs with a tennis ball in its mouth”



# Visual-language models are learned from image-caption pair like **this(?)**

---

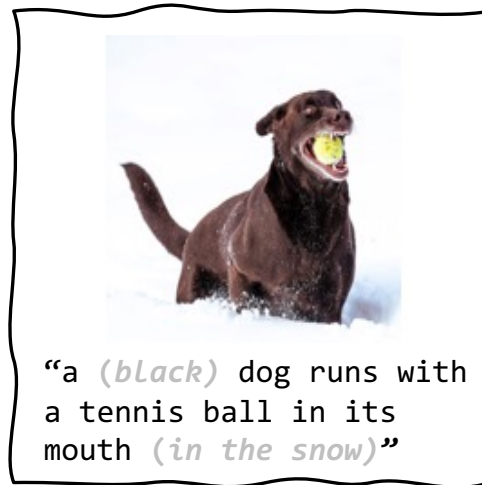
“A dog runs with a tennis ball in its mouth”



# Objects and attributes could be omitted in visual-language datasets

---

“A dog runs with a tennis ball in its mouth”

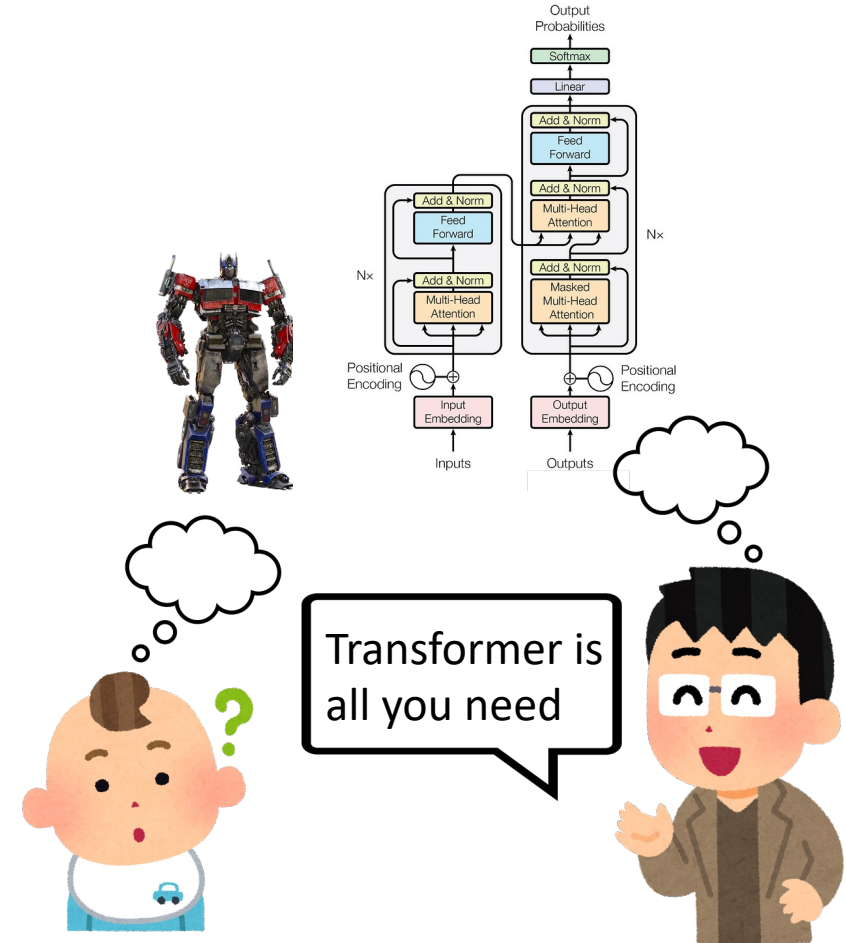


black ←.....→ brown  
snow ←.....→ grass



# *Reporting Bias* in visual-language datasets

- During human-human communications, people are prone to omit some attributes or properties of the topics they are describing:
  - They are assumed to be commonsense knowledge, so there is no need to report explicitly.



# *Reporting Bias* in visual-language datasets

---

- Current web-scale datasets are often collected (semi-) automatically from the human-human communication results from the Internet
  - E.g., image captions are created by someone, for someone, to convey some information
  - As a result, these datasets are not aware of the commonsense among people

# Do not assume the model has the same common knowledge like us

---



*"Common sense is not a gift, it's a punishment. Because you have to deal with everyone who doesn't have it."*

Salmon swimming in the river – from DALL·E 2

# Question?

---



# How *reporting bias* can affect visual-language learning?

---

- Biased captions
- Frequent object-attribute combination



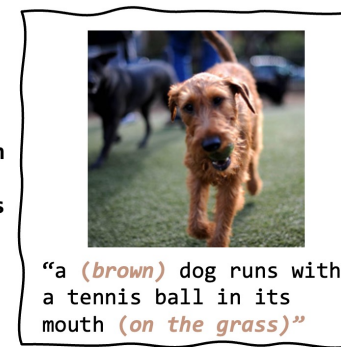
Biased captions with lacking objects or attributes, can be associated with dissimilar images.

Visual-language models do not naturally have the capability to grasp commonsense knowledge to discern the difference.

Ex.1. "A dog runs with a tennis ball in its mouth"



black  $\longleftrightarrow$  brown  
snow  $\longleftrightarrow$  grass



Ex.2. "Salmon dish on the table"



slices  $\longleftrightarrow$  fish





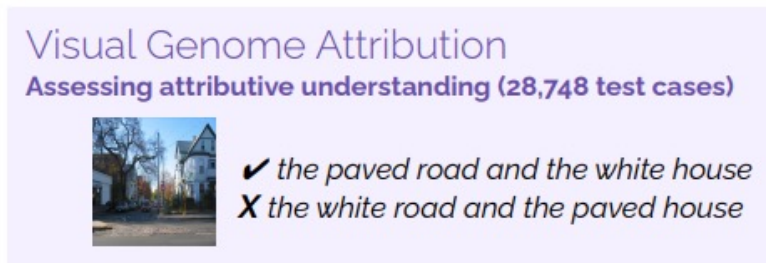
# Dominating frequent object-attribute combination

Learning from frequently occurring patterns hinders the model to distinguishing nuanced object-attribute combinations

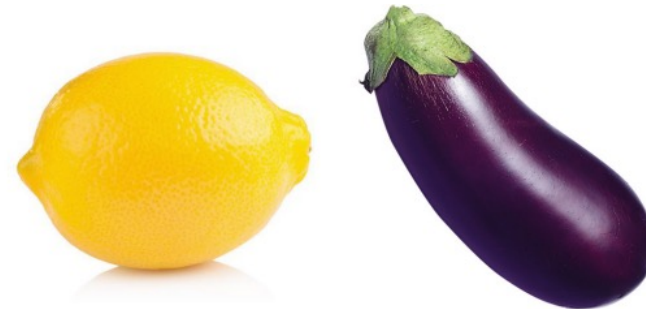


# One way to mitigate *reporting bias*: decoupling object-attribute association

- Reporting bias leads to aforementioned issues in the datasets, consequentially, the models trained on these datasets
  - CLIP-like models tend to be insensitive to the object-attribute association
  - E.g. The text of “lemon” is similar to the visual concept “yellow” in embedding space



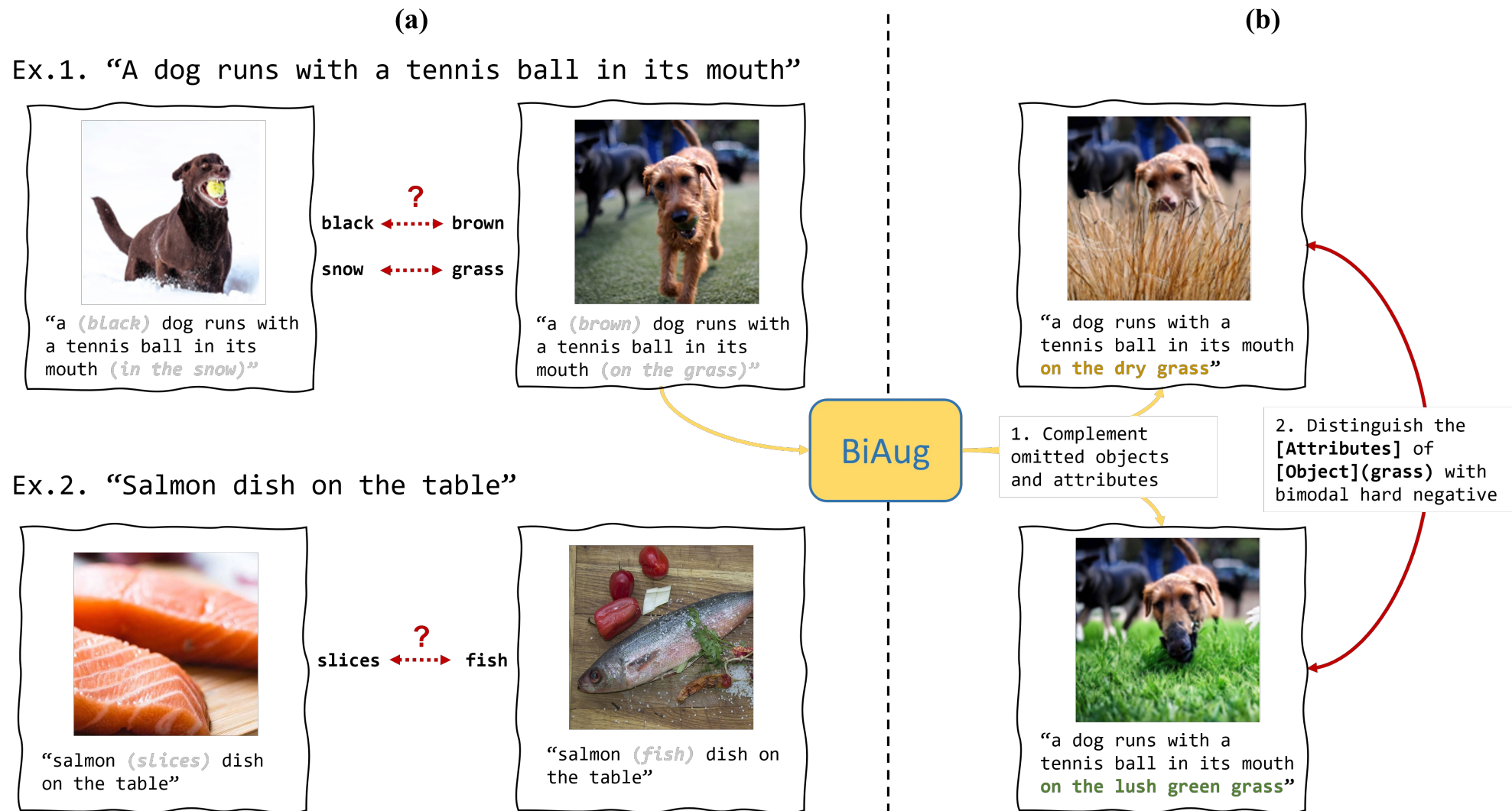
*Yuksekgonul et. al. ICLR2023*



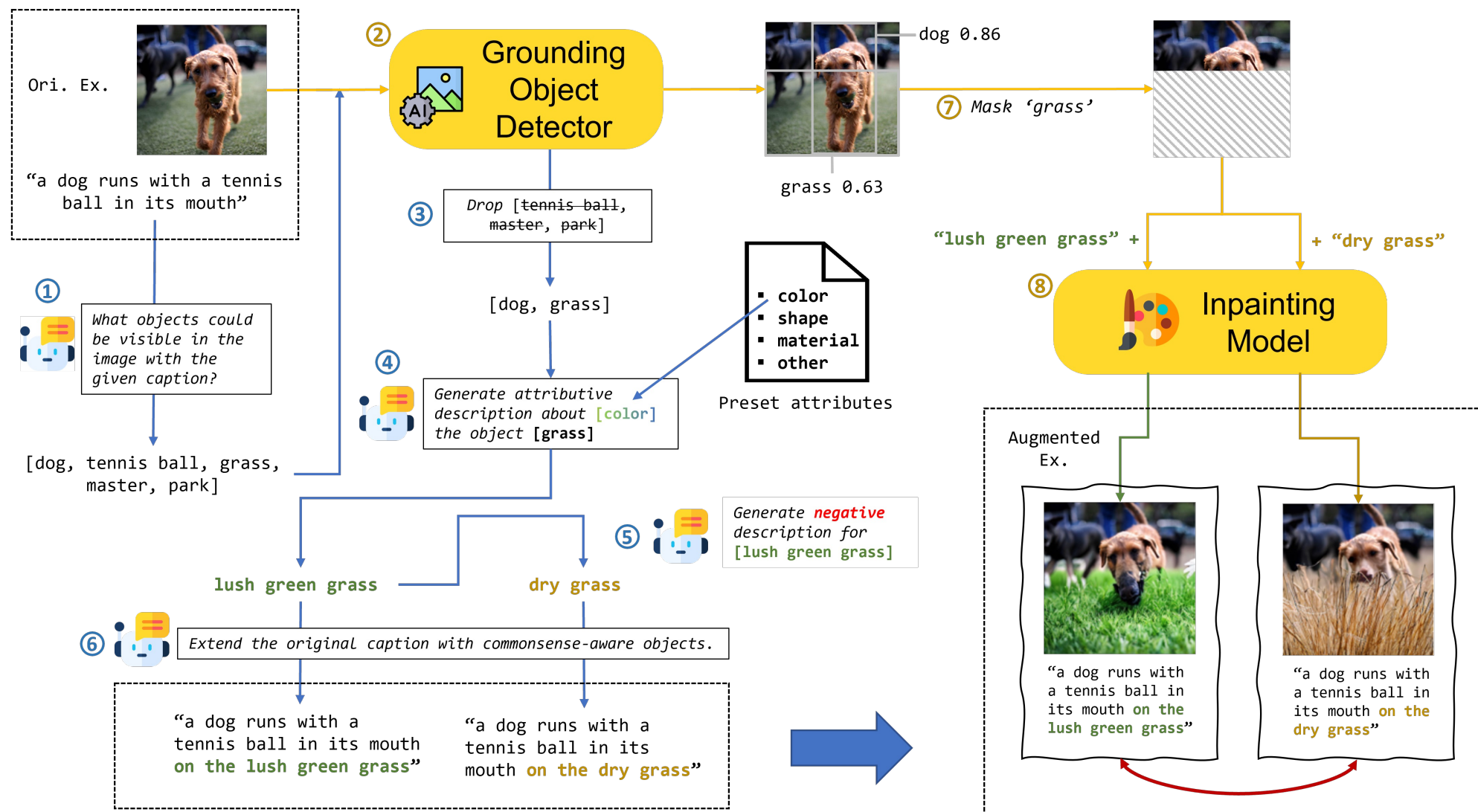
CLIP: "In this picture, the color of the lemon is purple."

*Yamada et. al. 2022*

# BiAug: Bimodal Augmentation by decoupling object-attribute association

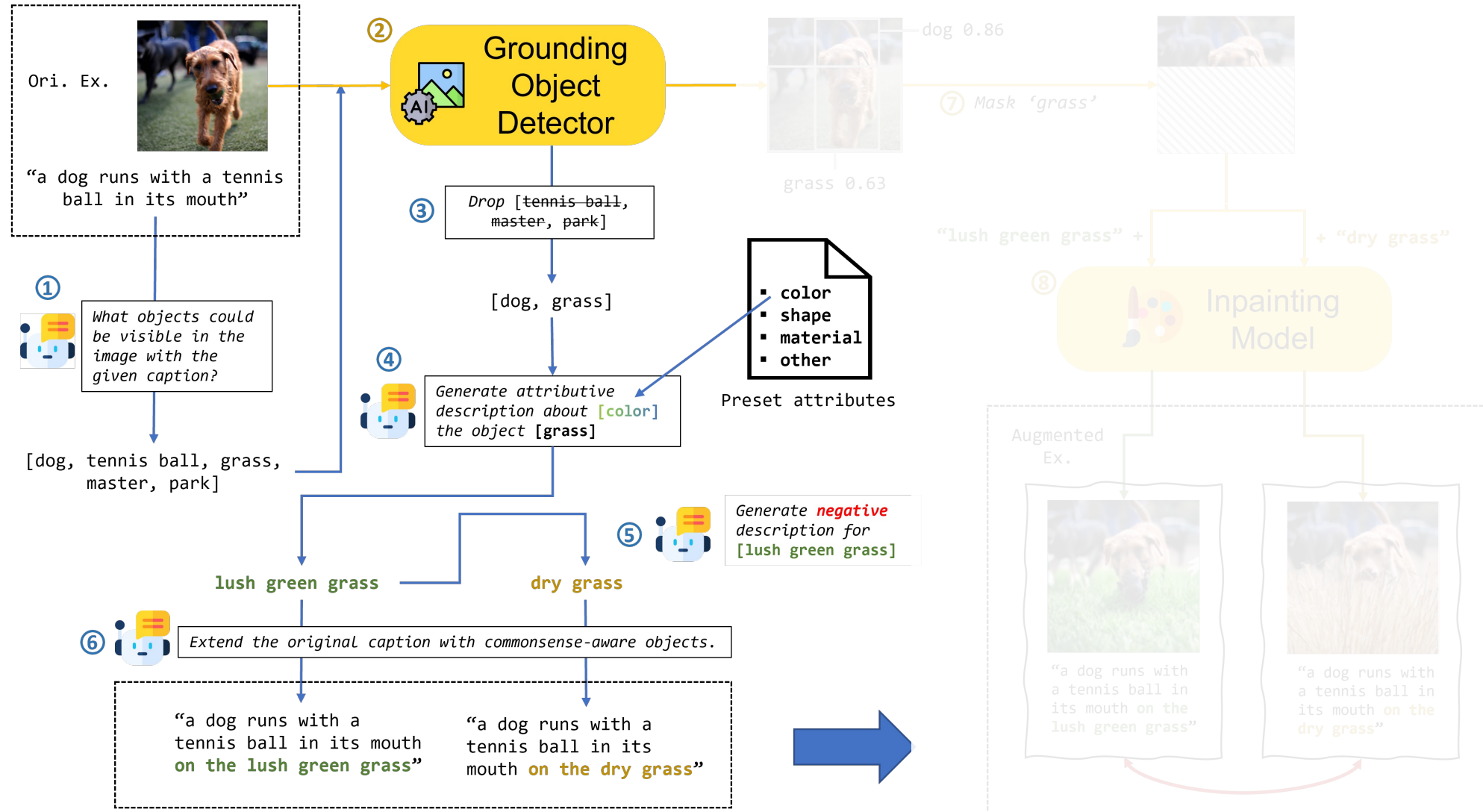


# Pipeline of BiAug

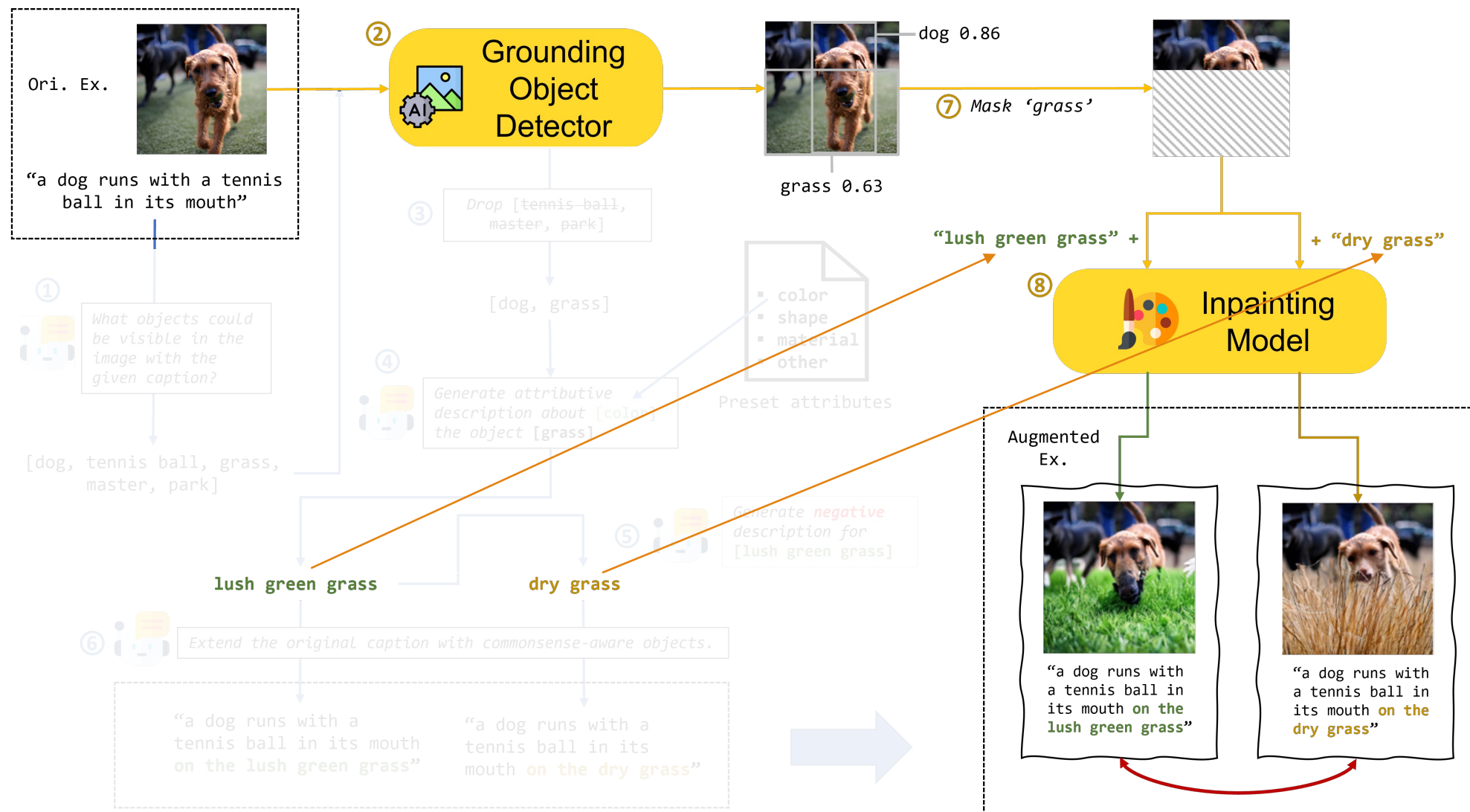




# Modality of language: utilize LLM to add explicit knowledge

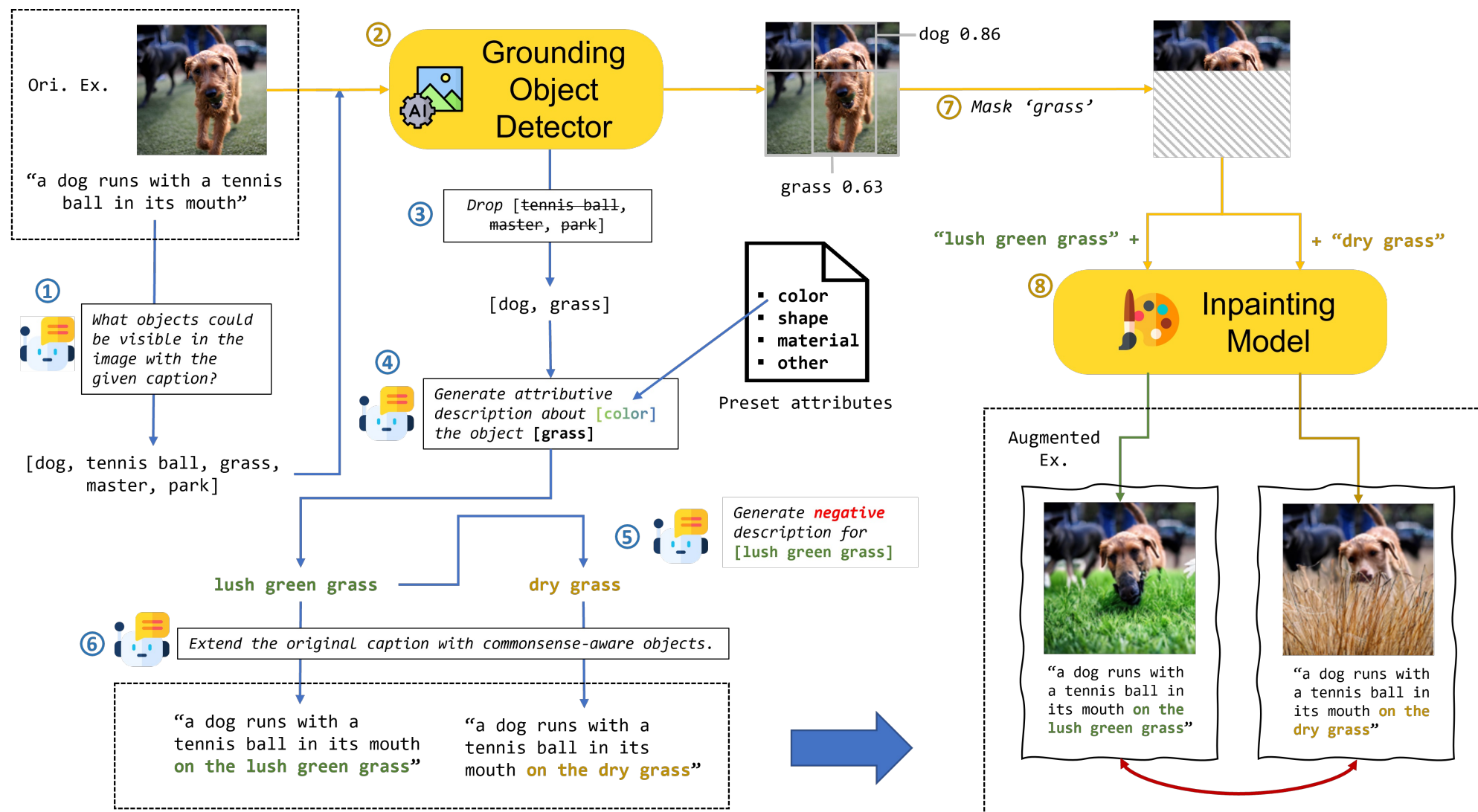


# Modality of vision: modify the object with inpainting model











# Pipeline of BiAug



# Question?

---

Ex.

	Ori.	Aug. 1	Aug. 2		Ori.	Aug. 1	Aug. 2
							
Ori.							
Aug. 1	“riders		were closely packed for the majority of the race”		“image result for beginners paintings of		boats and the sea”
	“riders	<u>on yellow bicycles</u>	were closely packed for the majority of the race”		“image result for beginners paintings of	<u>a blue</u>	boat and the sea”
Aug. 2	“riders	<u>on red bicycles</u>	were closely packed for the majority of the race”		“image result for beginners paintings of	<u>a red</u>	boat and the sea”
							
Ori.							
Aug. 1	“row of hay bales in a		field”		“		tractor and plough in an arable field”
	“a row of hay bales in a	<u>grassy</u>	field”		“	<u>a wooden</u>	tractor in an arable field”
Aug. 2	“a row of hay bales in a	<u>muddy</u>	field”		“	<u>a metal</u>	tractor in an arable field”
							
Ori.							
Aug. 1	“		large greenhouses at a country garden in summertime”		“use an		empty chair as a symbol of those missing from our lives on all saints day”
	“	<u>rectangular</u>	greenhouses at a country garden in summertime”		“use a	<u>broken</u>	empty chair as a symbol of those missing from our lives on all saints day”
Aug. 2	“	<u>circular</u>	greenhouses at a country garden in summertime”		“use a	<u>intact</u>	empty chair as a symbol of those missing from our lives on all saints day”

# Dataset construction

---

We extracted subsets of 40,000, 100,000, 200,000 and 300,000 examples from the Conceptual Caption 3M (CC3M) dataset, labeled as 40K, 100K, 200K and 300K respectively.

Table 1: Statistics of synthesized dataset. \*: some of the examples in the source dataset are dropped due to issues such as overly long sequence.

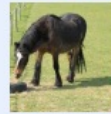
Source Dataset	40K	100K	200K	300K
# of source data *	38,100	88,300	187,900	287,600
# of extract objects	39,640	91,472	194,571	297,567
# of augmented examples	122,026	280,764	599,860	921,874
– after filtering	77,700	178,746	381,275	586,278
# of hard negative pairs	61,013	140,376	299,910	460,908
– after filtering	30,325	69,748	148,690	228,605

# Benchmark: Attribution, Relation, and Order benchmark (ARO) *Yuksekgonul et. al. ICLR2023*

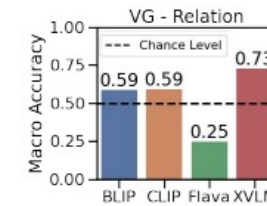
Fine-grained evaluation of vision language models' relation, attribution, and order understanding.

## Visual Genome Relation

Assessing relational understanding (23,937 test cases)



- ✓ the horse is eating the grass
- ✗ the grass is eating the horse

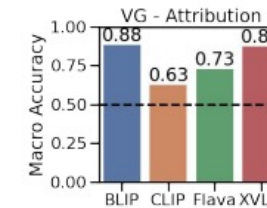


## Visual Genome Attribution

Assessing attributive understanding (28,748 test cases)



- ✓ the paved road and the white house
- ✗ the white road and the paved house

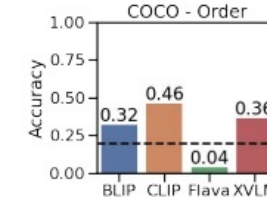
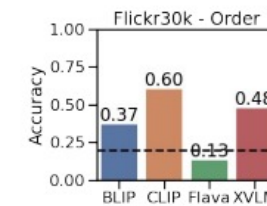


## COCO Order and Flickr Order

Assessing sensitivity to order (6,000 test cases)



- ✓ a brown cat is looking at a gray dog and sitting in a white bathtub
- ✗ (shuffle adjective/noun) a gray bathtub is looking at a white cat and sitting in a brown dog
- ✗ (shuffle all but adjective/noun) at brown cat a in looking a gray dog sitting is and a white bathtub
- ✗ (shuffle words within trigrams) cat brown a at is looking a gray dog in and sitting bathtub a white
- ✗ (shuffle trigrams) a brown cat a white bathtub is looking at a gray dog and sitting in



BLIP

the grass is eating the horse 81%

the horse is eating the grass 78%

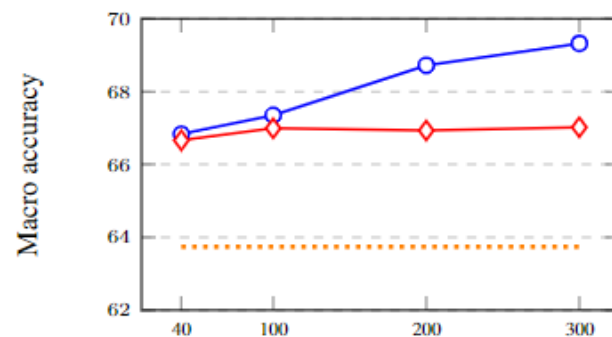


# Results on ARO

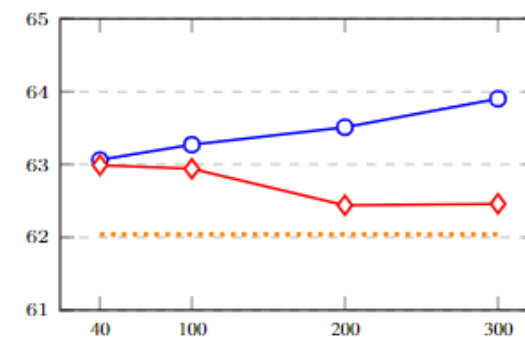
Yellow dotted line:  
original CLIP

Red line: CLIP ft w/  
source dataset

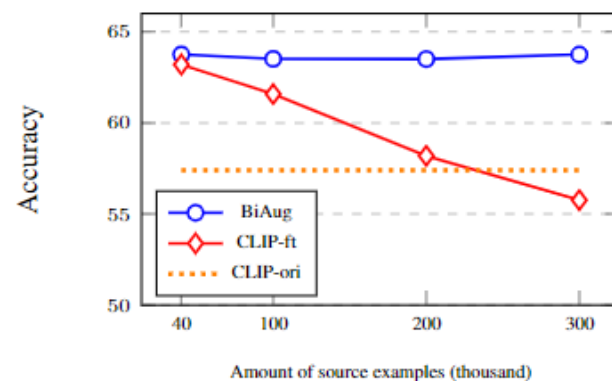
Blue line: CLIP ft w/  
BiAug dataset



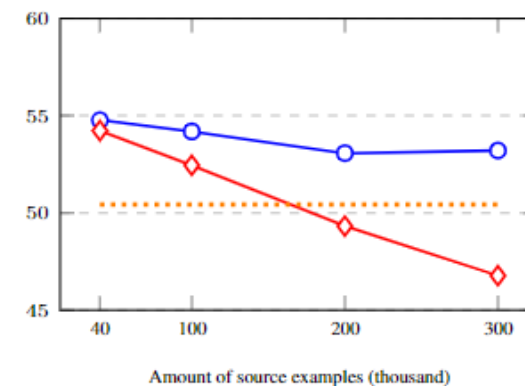
(a) VGenome-Relation



(b) VG-Attribute



(c) Flickr30K-Order



(d) COCO-Order

# Visual language retrieval

- We would also like to evaluate on general benchmark and verify no overfitting with small finetuning datasets
  - Reporting retrieval results on MSCOCO and Flickr30K
  - Image retrieval: input a caption and retrieve the corresponding image
  - Text retrieval: input an image and retrieve the correct caption(s)
    - Each image has 5 captions, such that text retrieval is simpler

Method	Image @1	Image @5	Image @10	Text @1	Text @5	Text @10
MSCOCO						
CLIP	33.07	58.41	68.98	52.38	76.72	84.60
CLIP-ft	34.59	59.84	70.23	54.78	78.08	85.34
BiAug	<b>35.60</b>	<b>60.57</b>	<b>70.68</b>	<b>55.46</b>	<b>78.78</b>	<b>86.20</b>
Flickr30K						
CLIP	62.08	85.58	91.78	81.90	96.20	<b>98.80</b>
CLIP-ft	64.72	86.94	92.04	82.00	97.00	98.70
BiAug	<b>65.36</b>	<b>87.26</b>	<b>92.76</b>	<b>82.20</b>	<b>97.10</b>	<b>98.80</b>

Table 2: Retrival results on MSCOCO and Flickr30K. Image @K denotes the image retrieval with recall@K. Text @K denotes the text retrieval recall@K.

# Summary and takeaways

---

1. Reporting bias exists broadly in web-crawled visual-language datasets.
2. Mitigating reporting bias could help visual-language model with two issues
  - Biased captions can compromise the training quality of VL models because they do not naturally have the **capability to grasp commonsense knowledge** to discern the difference.
  - Skewing the visual-language model towards **frequently occurring patterns**. This bias hinders the model's efficacy in distinguishing nuanced object–attribute combinations.
3. BiAug augments a source dataset with explicit commonsense knowledge from both visual and language modality.
  - Synthesize both new captions and corresponding images with explicit commonsense knowledge. BiAug crafts **bimodal hard negative examples** that emphasize a particular attribute.
  - Given that the object and attribute are decoupled, BiAug possesses the flexibility to produce samples with **a rich array of object–attribute pairings**.



# Thank you for your attention!

## Q&A

---

QIYU WU

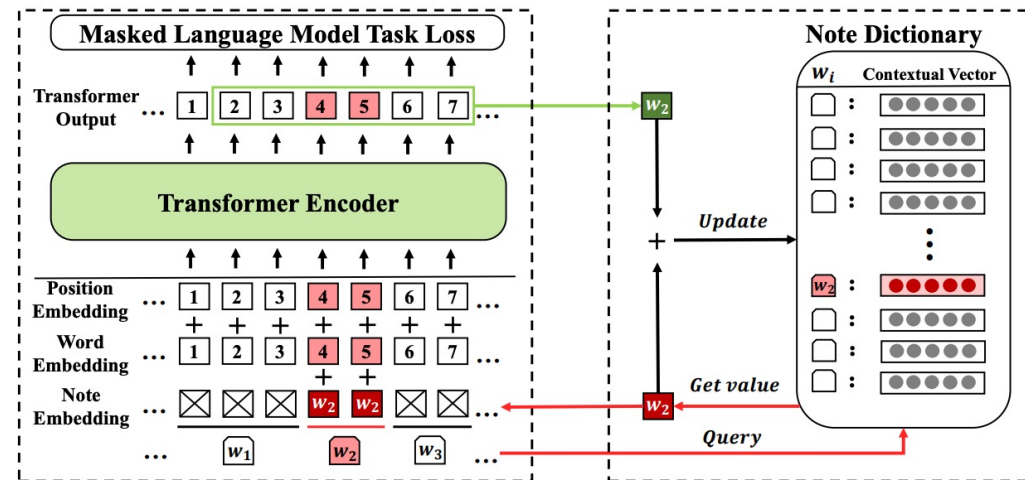
CONTACT: WUQIYU576@GMAIL.COM

# Appendix

---

PROJECTS I HAVE DONE IN THE PAST

# Taking notes on the fly helps language pre-training



Utilize cross-sentence signal to address rare words issue in language model pre-training([Wu et al. ICLR 2021](#))

# Rare words make inputs noisy, and slow down language training

Without Notes:

COVID-19 has cost thousands of \_\_\_\_\_ .

What is COVID-19?



dollars?  
donuts?  
puppies?  
tomatoes?

With Notes:

COVID-19 has cost thousands of lives .



Pandemic;  
global crisis

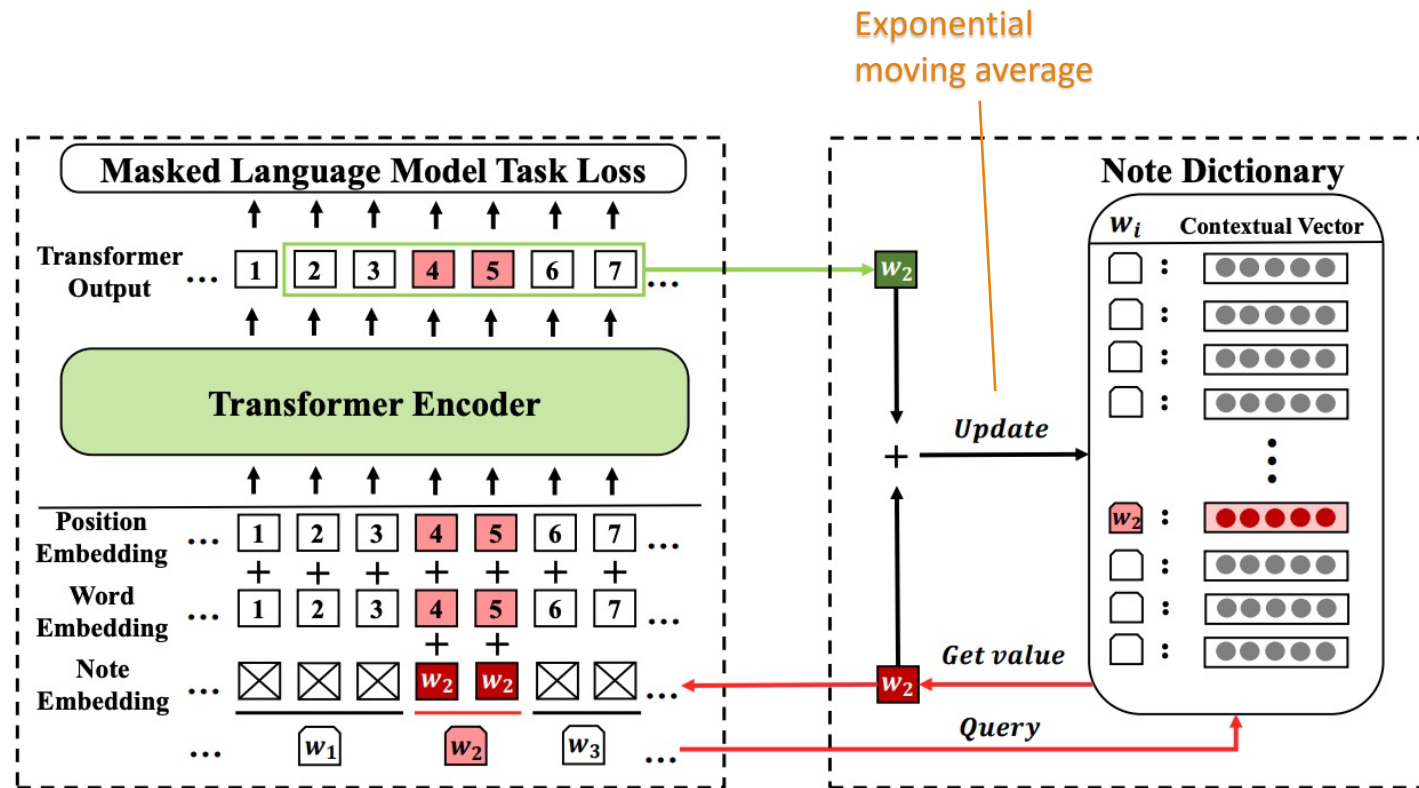
A note of 'COVID-19' taken from a previously seen sentence:

*The COVID-19 pandemic is an ongoing global crisis.*

Note-taking is a useful skill which can help people recall information that would otherwise be lost.

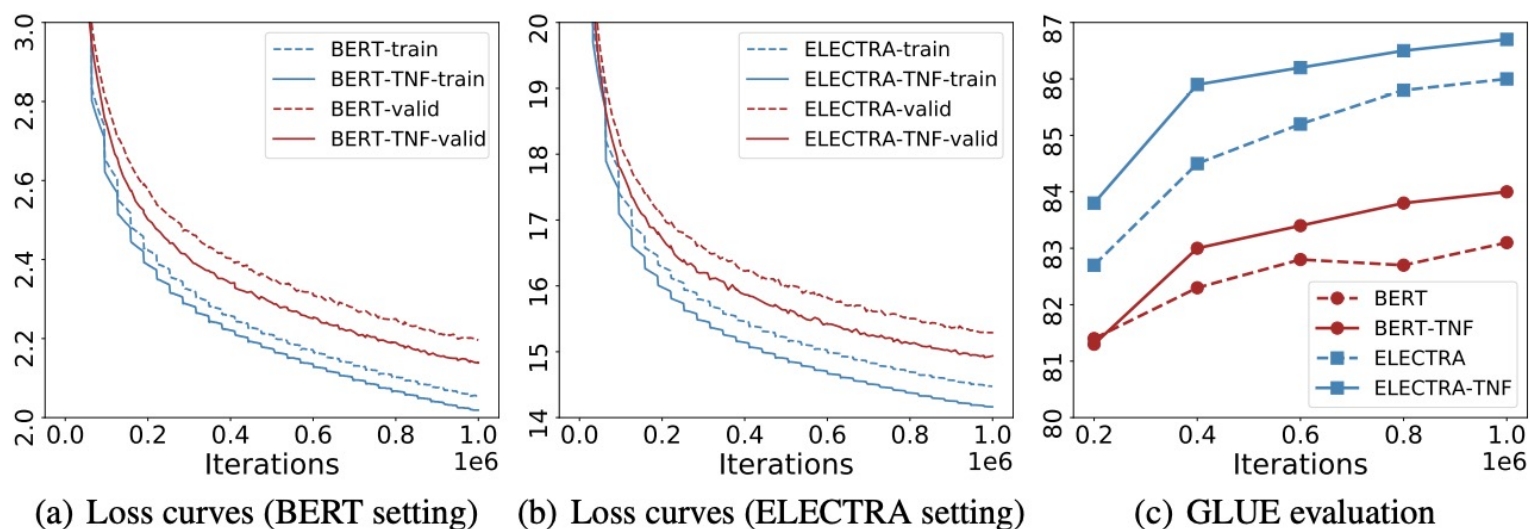
In our dataset (Wikipedia and BookCorpus containing 3.47B words), 20% of sentences and 90% of inputs contain at least one rare word (200K with frequency 100 - 500).

# Taking notes helps language pre-training



$w_2$  is a rare word

# Taking notes expedites language pre-training



Save 60%  
pretraining time!

Figure 3: The curves of pre-training loss, pre-training validation loss and average GLUE score for all models trained under the BERT setting and ELECTRA setting. All three sub-figures show that TNF expedites the backbone methods.

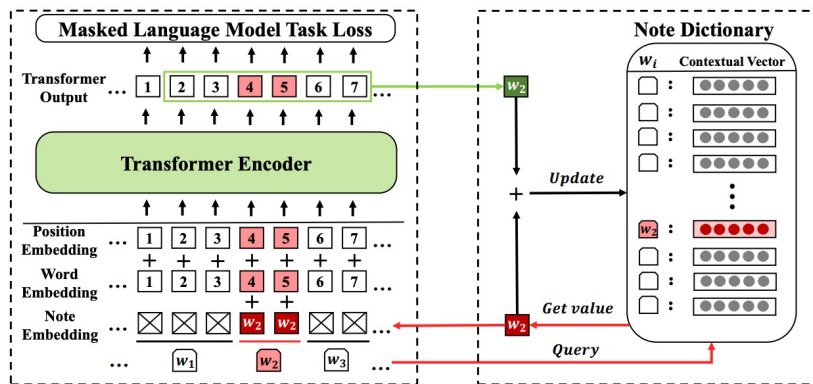
# Note dictionary can be removed after pre-training is finished

	MNLI	QNLI	QQP	SST	CoLA	MRPC	RTE	STS	Avg.
BERT (Ours)	85.0	<b>91.5</b>	91.2	93.3	58.3	88.3	69.0	88.5	83.1
BERT-TNF	85.0	91.0	<b>91.2</b>	93.2	59.5	<b>89.3</b>	<b>73.2</b>	<b>88.5</b>	<b>83.9</b>
BERT-TNF-F	<b>85.1</b>	90.8	91.1	93.3	59.8	88.8	72.1	88.5	83.7
BERT-TNF-U	85.0	90.9	91.1	<b>93.4</b>	<b>60.2</b>	88.7	71.4	88.4	83.6
ELECTRA(Ours)	86.8	92.7	91.7	93.2	66.2	<b>90.2</b>	76.4	<b>90.5</b>	86.0
ELECTRA-TNF	<b>87.0</b>	<b>92.7</b>	<b>91.8</b>	93.6	<b>67.0</b>	90.1	81.2	90.1	<b>86.7</b>
ELECTRA-TNF-F	86.9	92.6	91.8	<b>93.7</b>	65.9	89.7	<b>81.4</b>	89.8	86.5
ELECTRA-TNF-U	86.9	92.7	91.7	93.6	66.3	89.8	81.0	89.8	86.5

Table 2: Performance of different models on downstream tasks. Results show that TNF outperforms backbone methods on the majority of individual tasks. We also list the performance of two variants of TNF. Both of them leverage the node dictionary during fine-tuning. Specifically, TNF-F uses fixed note dictionary and TNF-U updates the note dictionary as in pre-training. Both models outperforms the baseline model while perform slightly worse than TNF.

# Takeaways

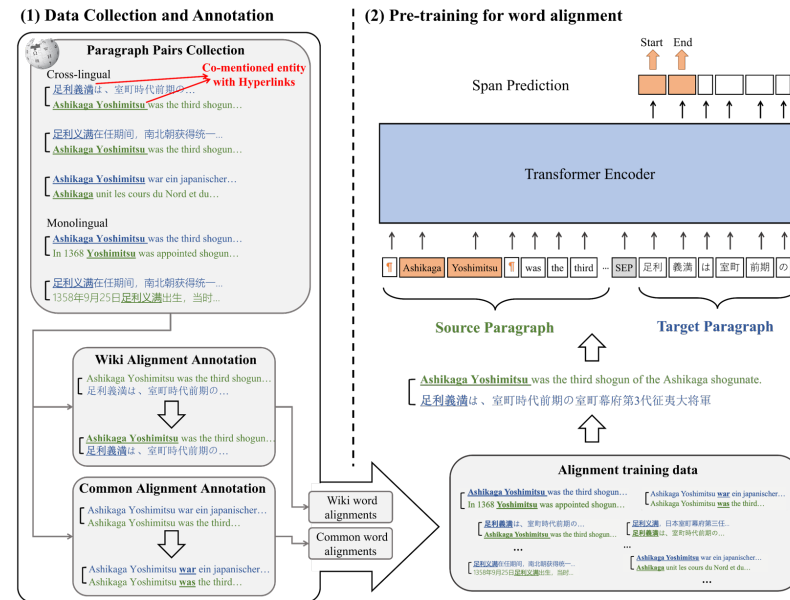
1. Rare words make input noisy, which can slow down optimization of the whole model.
2. Taking notes during the pre-training can outperform baselines on GLUE with 40% pre-training time.
3. The note dictionary can be removed after the pre-training is finished.



Utilize cross-sentence signal to address rare words issue in language model pre-training([Wu et al. ICLR 2021](#))



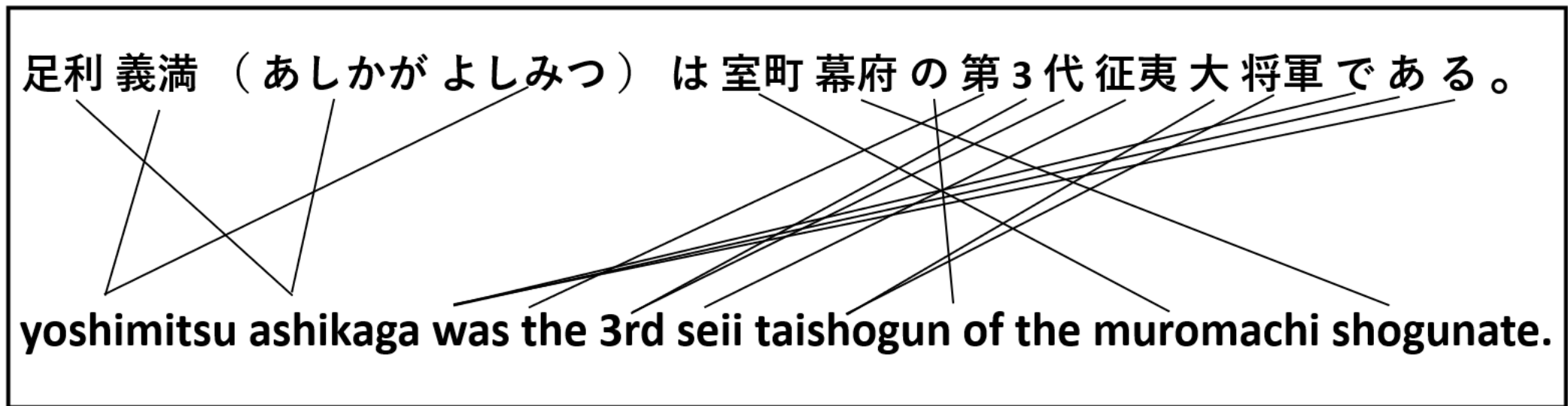
# Weakly supervised word alignment pre-training



Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training ( [Wu et al. ACL 2023](#) )

# Word Alignment

word alignment aims to align the corresponding words in parallel texts.



# Do we really need manual alignment data to do word alignment?

---

Most existing word alignment methods rely on either **manual alignment datasets or parallel corpora** for training, which weakens their usefulness because of the limiting accessibility of data.

We relax the requirements for:

- **correct** (manually made),
- **fully-aligned** (all words in a sentence pair are annotated),
- **parallel sentences**.

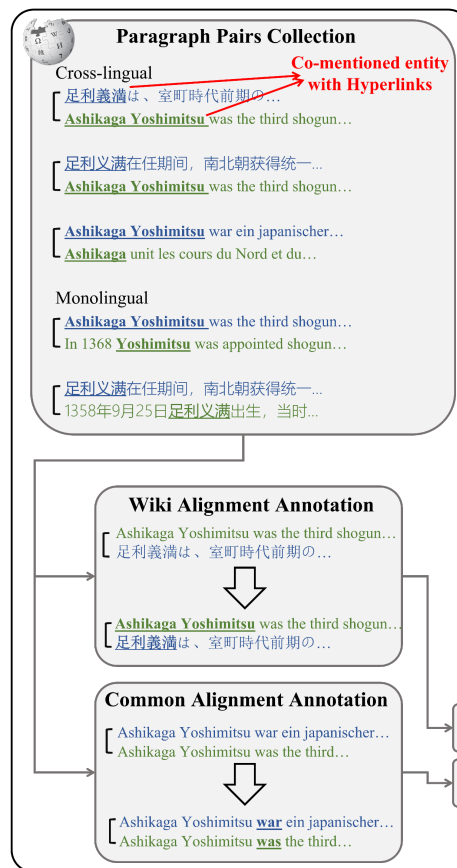
Specifically, we make **a large-scale (2 million pairs) training data** that are:

- **noisy (automatically made)**,
- **partially-aligned**,
- **non-parallel paragraphs** (or mono-lingual paragraph pairs).

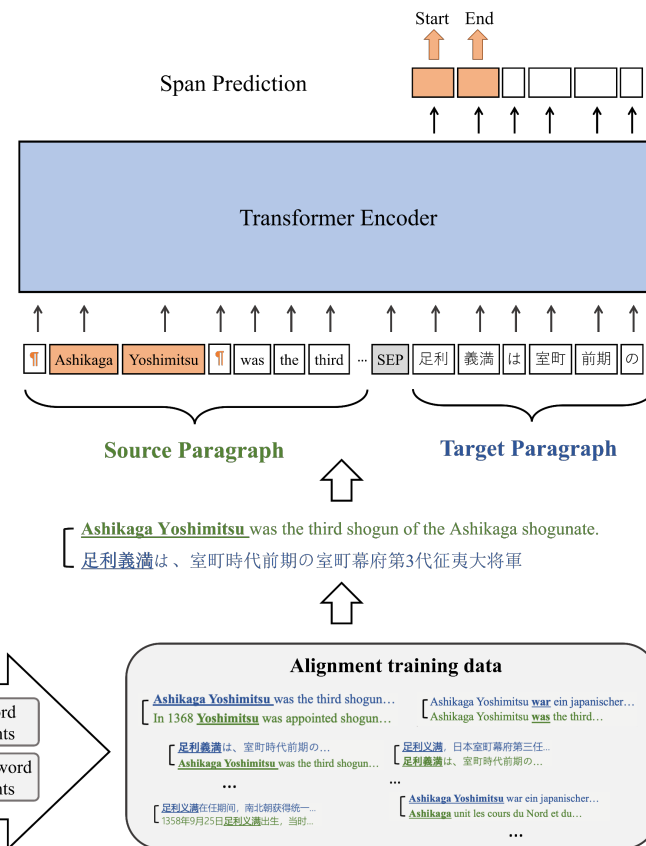
# Approach: word alignment pre-training via large-scale weakly supervised span prediction

- Data Collection
- Common word annotation
- Wiki word Annotation
- Span-prediction Pre-training

## (1) Data Collection and Annotation



## (2) Pre-training for word alignment



# Paragraph pair collection

## (1) Data Collection and Annotation

- Data Collection
- Common word annotation
- Wiki word Annotation
- Span-prediction Pre-training



### Paragraph Pairs Collection

Cross-lingual

[ [足利義満](#)は、室町時代前期の...  
[Ashikaga Yoshimitsu](#) was the third shogun...

[ [足利義満](#)在任期间，南北朝获得统一...  
[Ashikaga Yoshimitsu](#) was the third shogun...

[ [Ashikaga Yoshimitsu](#) war ein japanischer...  
[Ashikaga](#) unit les cours du Nord et du...

Monolingual

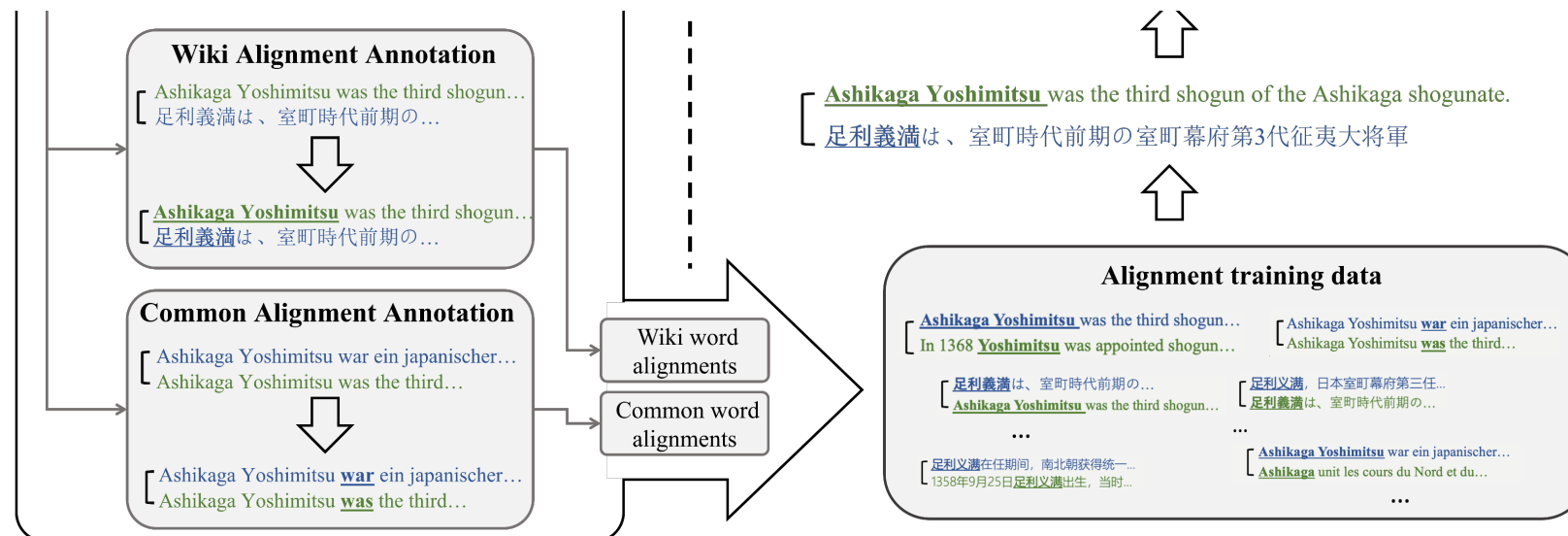
[ [Ashikaga Yoshimitsu](#) was the third shogun...  
In 1368 [Yoshimitsu](#) was appointed shogun...

[ [足利義満](#)在任期间，南北朝获得统一...  
1358年9月25日[足利義満](#)出生，当时...

Collect both mono-lingual and Cross-lingual Wikipedia paragraph pairs by **co-mentioned hyperlinks**.

# Alignment annotation

- Data Collection
- Common word annotation
- Wiki word Annotation
- Span-prediction Pre-training



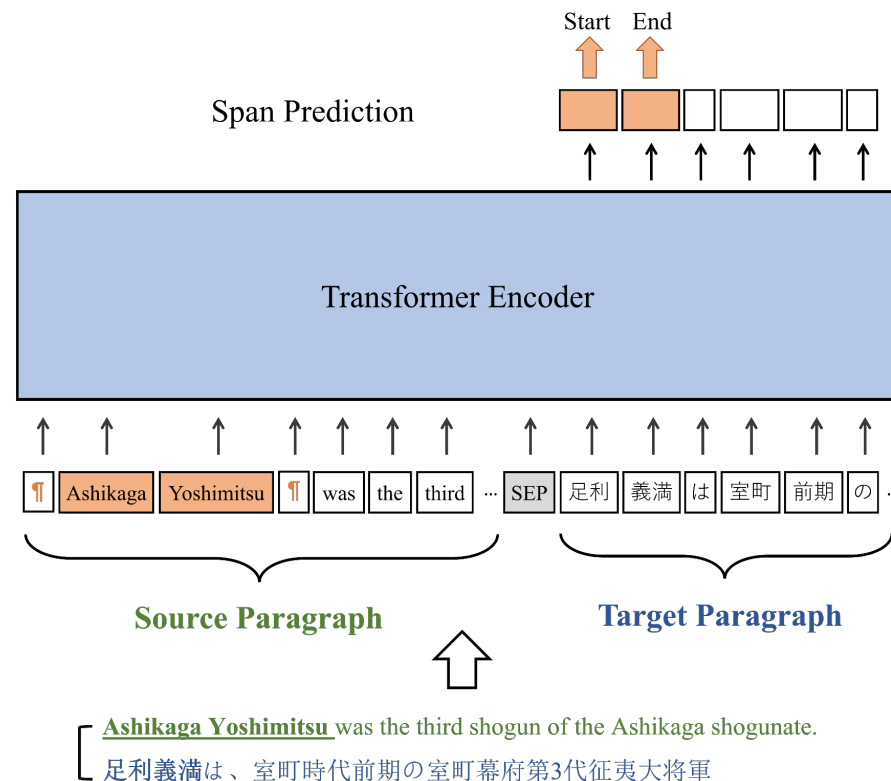
- **Make Common word annotation** by bi-directional agreement, with contextual embeddings in a pre-trained language model.

- **Make Wiki word Annotation** by directly aligning the corresponding hyperlinks spans of the co-mentioned entity.

# Span prediction pre-training

## (2) Pre-training for word alignment

- Data Collection
- Common word annotation
- Wiki word Annotation
- **Span-prediction Pre-training**



- Given a source paragraph with a source token specified by the **special token** ¶, the goal is to predict the aligned tokens in the target paragraph.
- Concatenate the source and target paragraph as input sequence and perform the **span prediction task**.

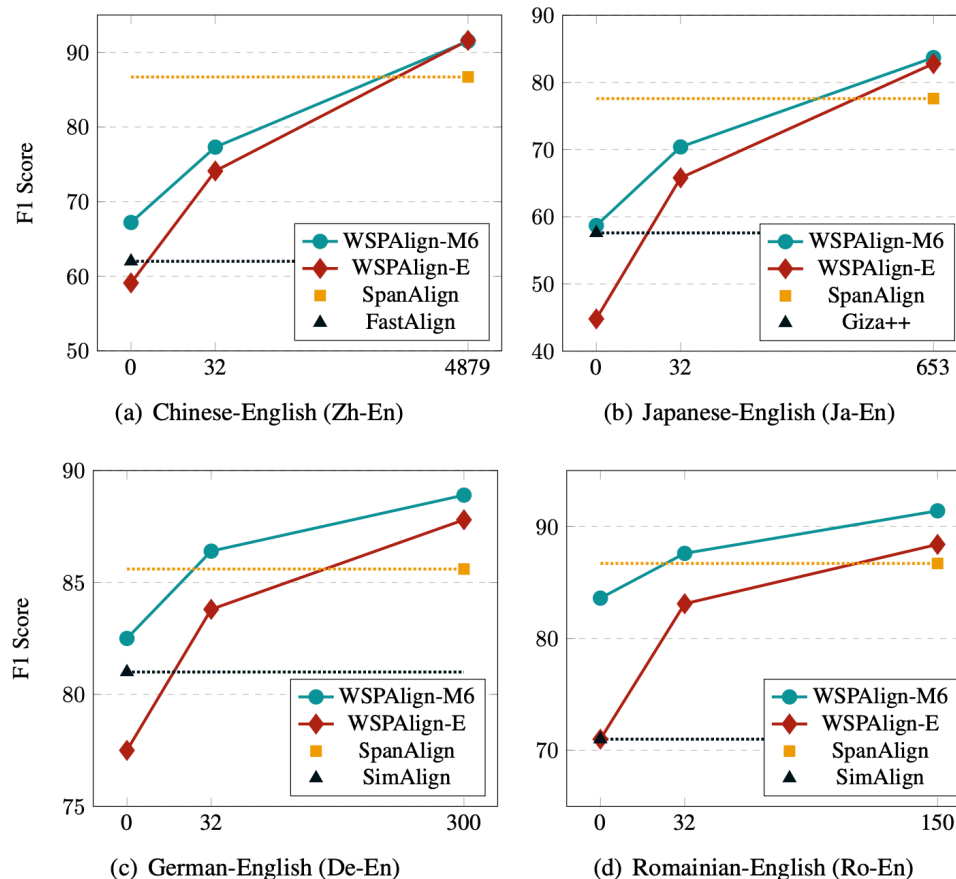
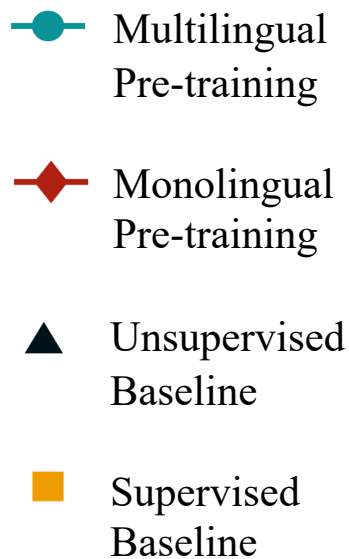
# Experiments

Test Set	Method	Precision	Recall	F1	AER
Zh-En	FastAlign (Stengel-Eskin et al.)	80.5	50.5	62.0	-
	DiscAlign (Stengel-Eskin et al.)	72.9	74.0	73.4	-
	SpanAlign (Nagata et al., 2020)	84.4	89.2	86.7	13.3
	WSPAlign (ours)	90.8	92.2	<b>91.5</b> (↑ 4.8)	<b>8.5</b> (↓ 4.8)
Ja-En	Giza++ (Neubig, 2011)	59.5	55.6	57.6	42.4
	AWESoME (Dou and Neubig, 2021)	-	-	-	37.4
	SpanAlign (Nagata et al., 2020)	77.3	78.0	77.6	22.4
	WSPAlign (ours)	81.6	85.9	<b>83.7</b> (↑ 6.1)	<b>16.3</b> (↓ 6.1)
De-En	SimAlign (Jalili Sabet et al., 2020)	-	-	81.0	19.0
	AWESoME (Dou and Neubig, 2021)	-	-	-	15.0
	SpanAlign (Nagata et al., 2020)	89.9	81.7	85.6	14.4
	WSPAlign (ours)	90.7	87.1	<b>88.9</b> (↑ 3.3)	<b>11.1</b> (↓ 3.3)
Ro-En	SimAlign (Jalili Sabet et al., 2020)	-	-	71.0	29.0
	AWESoME (Dou and Neubig, 2021)	-	-	-	20.8
	SpanAlign (Nagata et al., 2020)	90.4	85.3	86.7	12.2
	WSPAlign (ours)	92.0	90.9	<b>91.4</b> (↑ 4.7)	<b>8.6</b> (↓ 3.6)
En-Fr	SimAlign (Jalili Sabet et al., 2020)	-	-	93.0	7.0
	AWESoME (Dou and Neubig, 2021)	-	-	-	4.1
	SpanAlign (Nagata et al., 2020)	97.7	93.9	-	4.0
	WSPAlign (ours)	98.8	96.0	-	<b>2.5</b> (↓ 1.5)

Table 1: Comparison of WSPAlign and previous methods on word alignment datasets. Higher F1 scores are better. Lower AER scores are better. We highlight the best number in the same setting and test set with bold font.



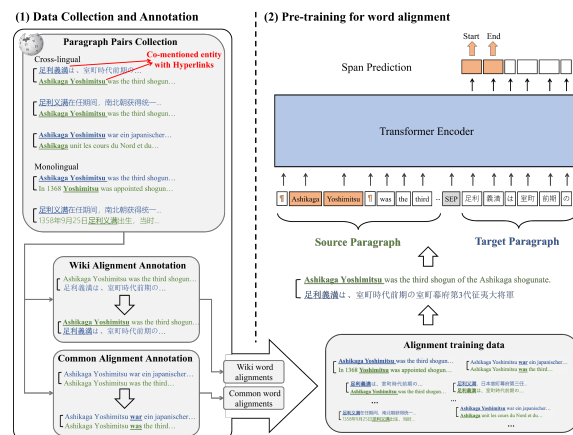
# Few-shot, fine-tuning and mono-lingual pre-training



- WSPAlign can be significantly improved and outperforms the existing unsupervised baselines with **few-shot** examples, which can be collected at a low cost.
- If we further fine-tune WSPAlign with a **full supervised dataset**, it can outperform the supervised baseline on all test sets.
- The improvement holds for **mono-lingual pre-training**.

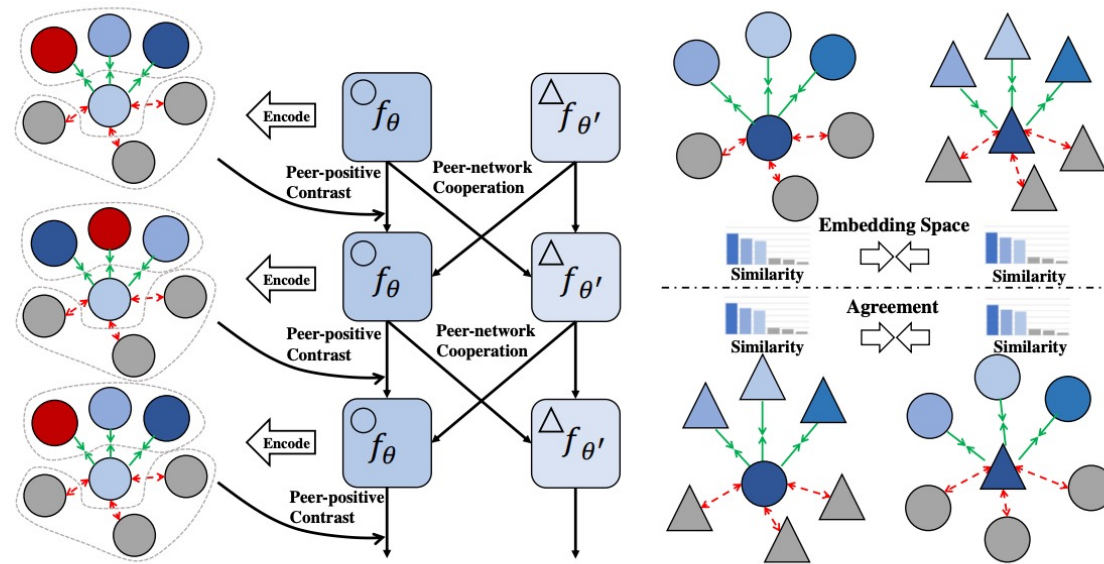
# Takeaways

1. We don't have to make perfect (correct, fully-aligned, parallel corpus) datasets to train word aligner.
2. Instead, weak supervision in **large-scaled** unlabeled text (noisy, partial, non-parallel) can be utilized for pre-training.
3. Zero-shot WSPAlign can outperform unsupervised baseline; few-shot and full-shot finetuning can further improve it and outperform supervised baseline.
4. Mono-lingual pre-training can be transferred to cross-lingual evaluation.



Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training ( [Wu et al. ACL 2023](#) )

# Diverse augmentation for sentence embeddings



Unsupervised data augmentation for sentence embedding by contrastive learning( [Wu et al. EMNLP 2022](#))

# Biases in unsupervised sentence embedding with contrastive learning

---

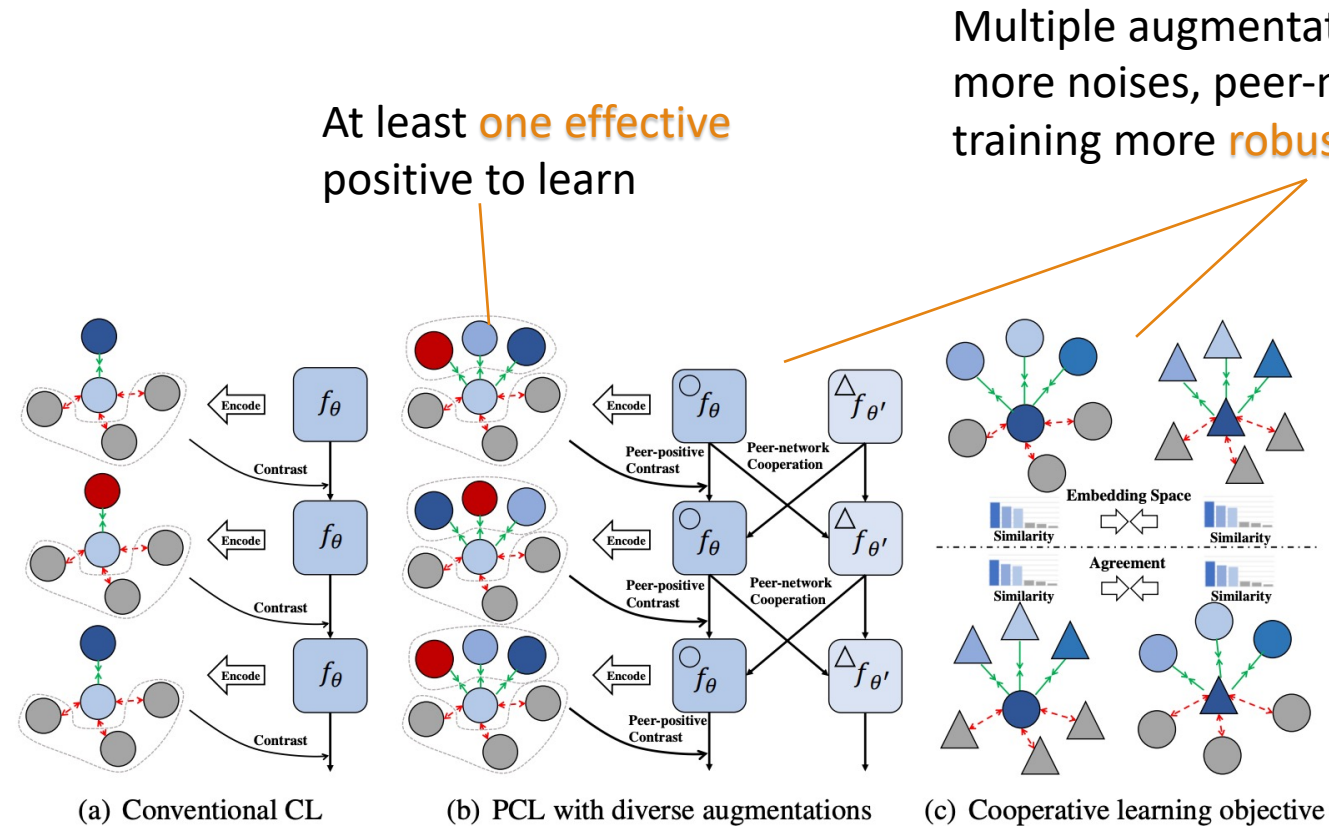
Contrastive learning is a common solution for sentence embedding

Data augmentation can construct positive instance.

However, text augmentation strategies change semantics in the sentence but still has shortcuts to learn.

Augmenting	Order	N-gram	Bag-of-words
<i>Shuffled Sentence</i>	×	×	✓
<i>Inversed Sentence</i>	×	✓	✓
<i>Word Repetition</i>	✓	×	✓
<i>Word Deletion</i>	✓	×	×

# Utilize **multiple** and **diverse** augmentations to construct training examples



# Both the number and diversity of augmentations are important

More augmentations is better!

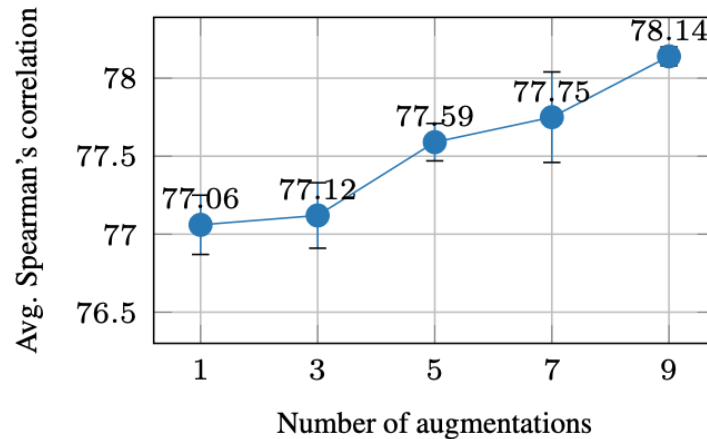


Figure 2: Effect of the number of augmentations.

Diverse augmentation is better!

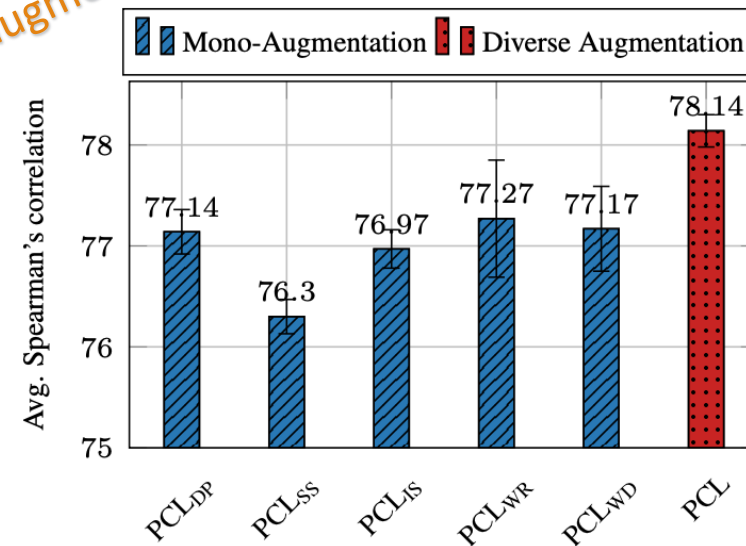
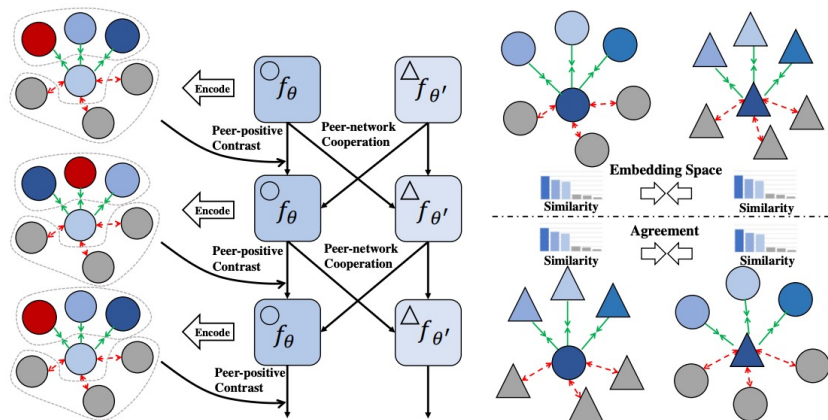


Figure 3: Effect of the diversity of augmentations.

# Takeaways

1. Single augmentation can be biased in contrastive learning for unsupervised sentence embeddings.
2. Utilizing **multiple** and **diverse** augmentation can mitigate the bias issue.
3. Dual networks can make the training with multiple positives more robust.



Unsupervised data augmentation for sentence embedding by contrastive learning([Wu et al. EMNLP 2022](#))

# Contact information

---

Name: Qiyu Wu

Linkedin: <https://www.linkedin.com/in/qiyuw/>

X/Twitter: @qiyuwwwv      *pls follow me :)*

Email: [wuqiyu576@gmail.com](mailto:wuqiyu576@gmail.com)

Address: Hongo Campus, The University of Tokyo, Bunkyo, Tokyo