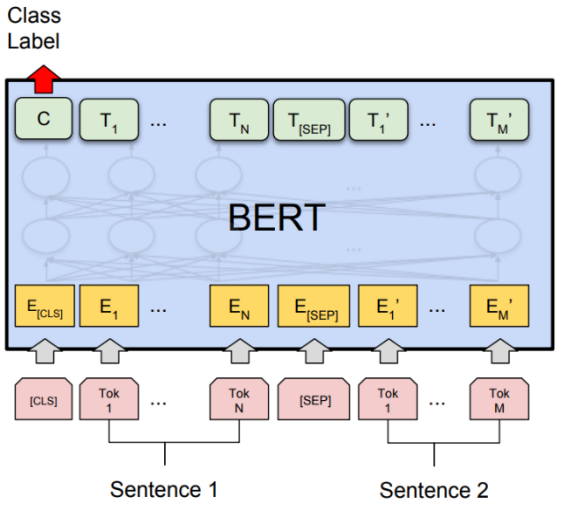
Vision-Language Pre-training: Progress and Challenge
视觉-语言预训练：进展和挑战

Nan DUAN (段楠)
Natural Langauge Computing Group
Microsoft Research Asia
2021

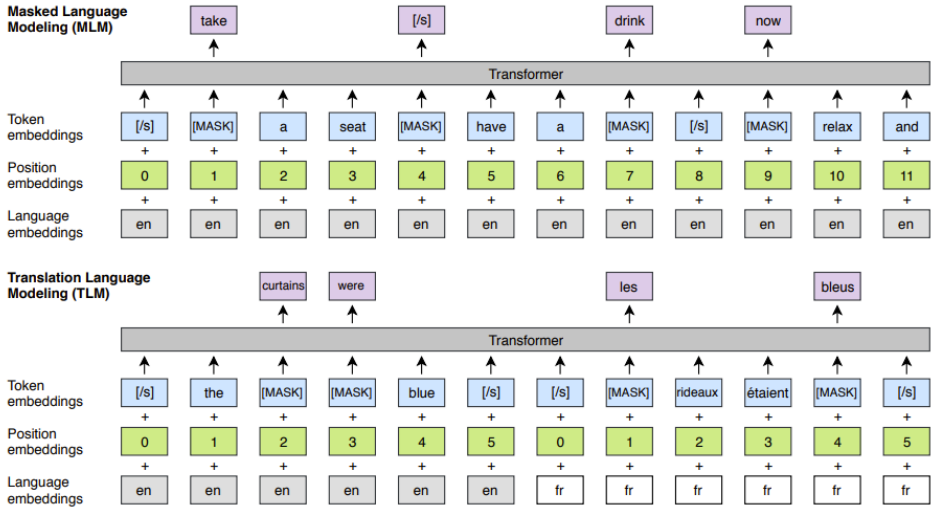
Outline

- Language Pre-training
- Vision Pre-training
- Vision-Language Pre-training
- Language-enhanced CV
- Vision-enhanced NLP
- Summarization

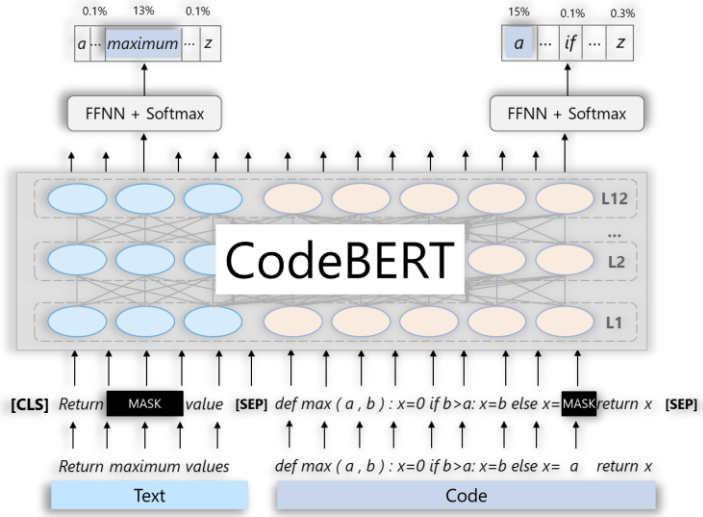
Current NLP Paradigm: Pre-trained Models



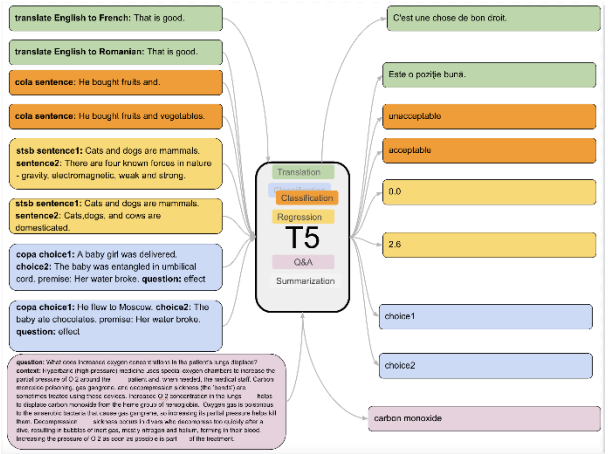
BERT (Devlin et al., 2018)



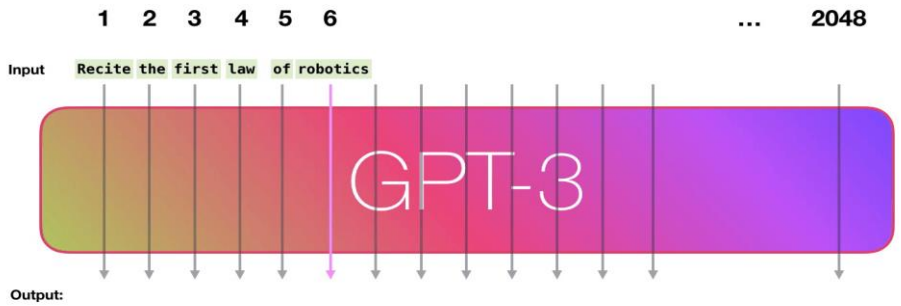
XLM (Lample and Conneau, 2019)



CodeBERT (Feng et al., 2020)



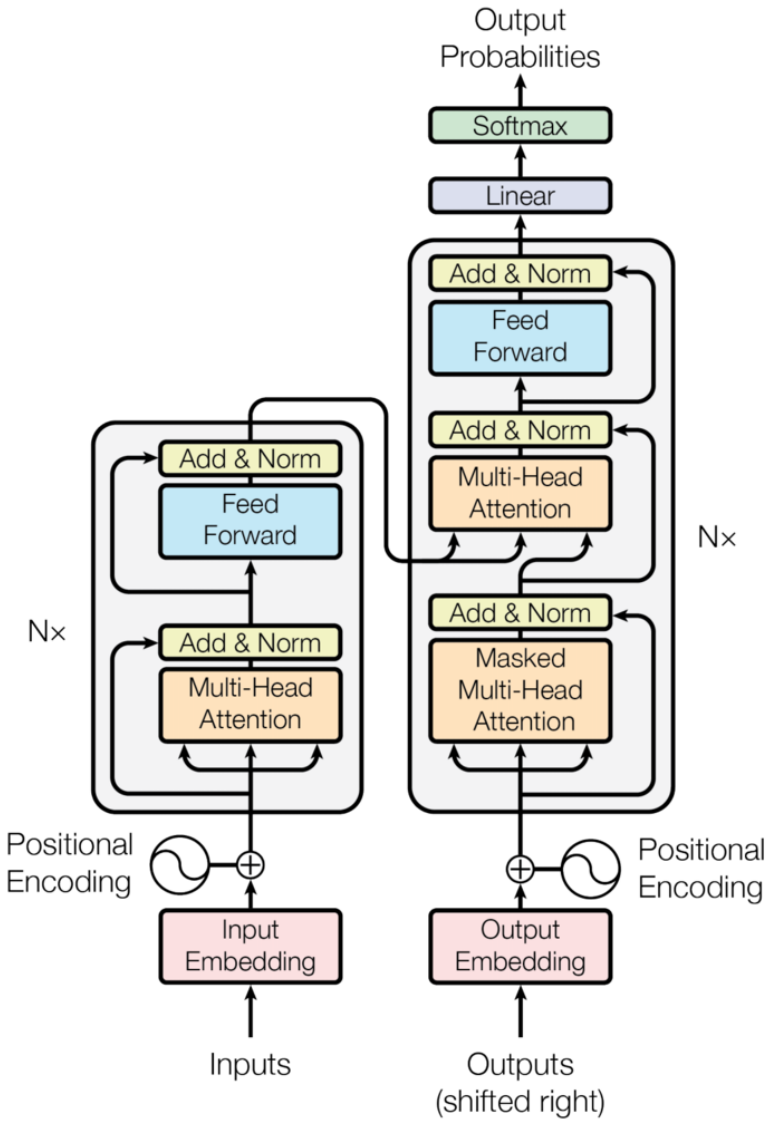
T5 (Raffel et al., 2020)



GPT-3 (Brown et al., 2020)

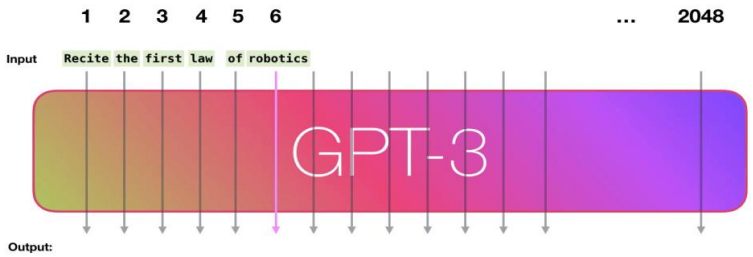
... ...

Transformer as Backbone



Self-supervised Learning as Pre-training

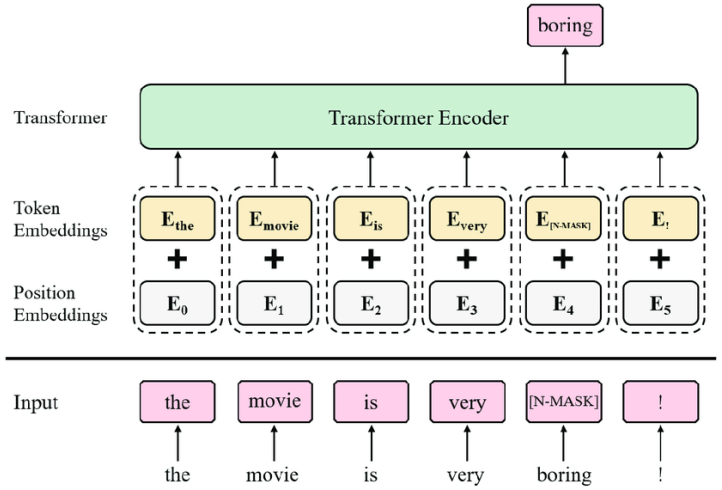
Auto-regressive Decoding



GPT-3 (Brown et al., 2020)

maximize the likelihood under the forward auto-regressive factorization

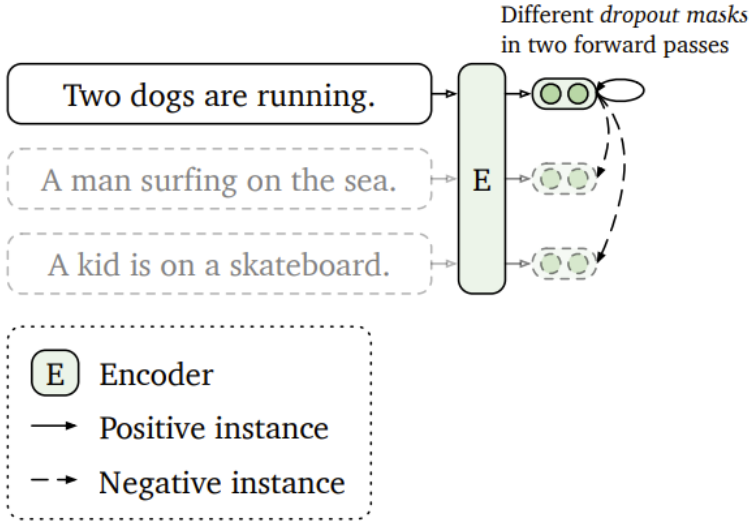
Denoising Auto-encoding



BERT (Devlin et al., 2018)

reconstruct original words/spans/sentences from corrupted sentences

Contrastive Learning



SimCSE (Gao et al., 2021)

learn representations such that similar samples stay close to each other

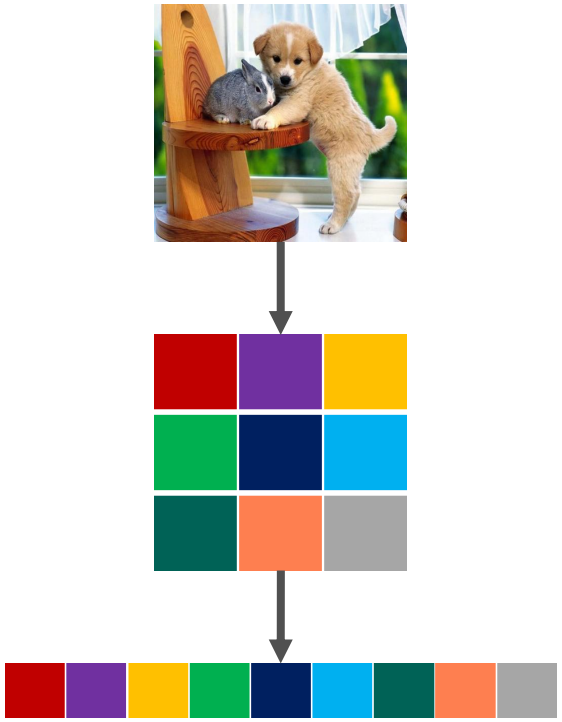
Outline

- Language Pre-training
- Vision Pre-training
- Vision-Language Pre-training
- Language-enhanced CV
- Vision-enhanced NLP
- Summarization

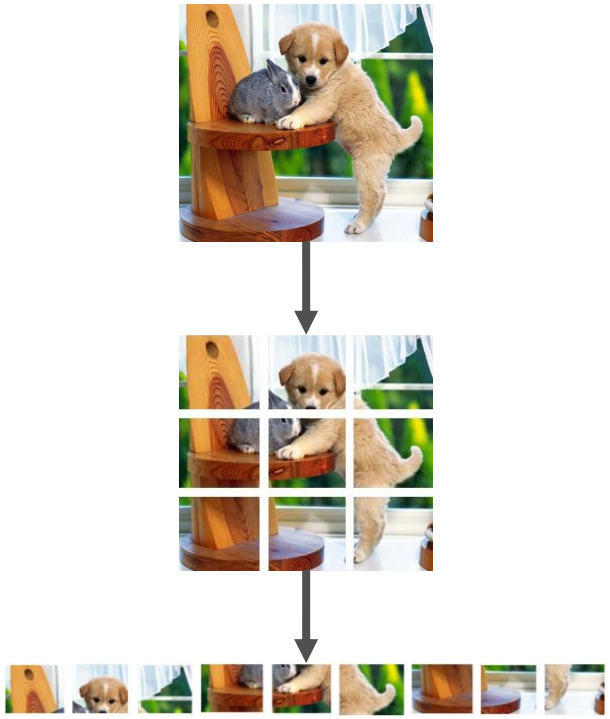
From Language Pre-training to Vision Pre-training

How to represent continuous vision contents in a discontinuous way?

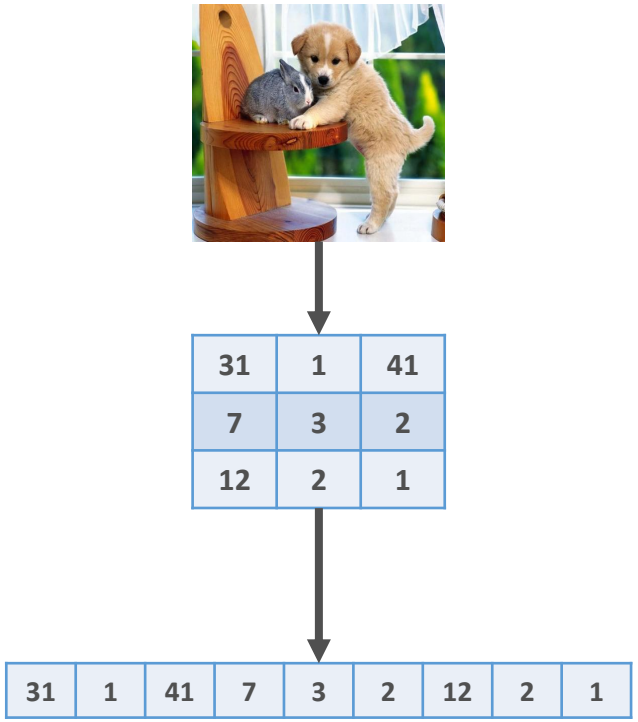
Pixel



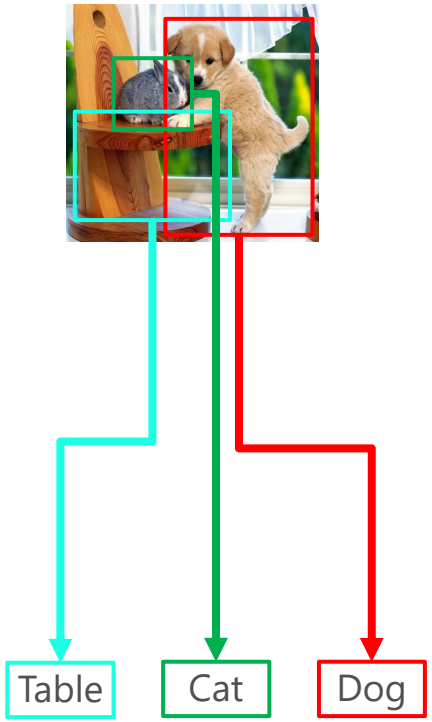
Patch



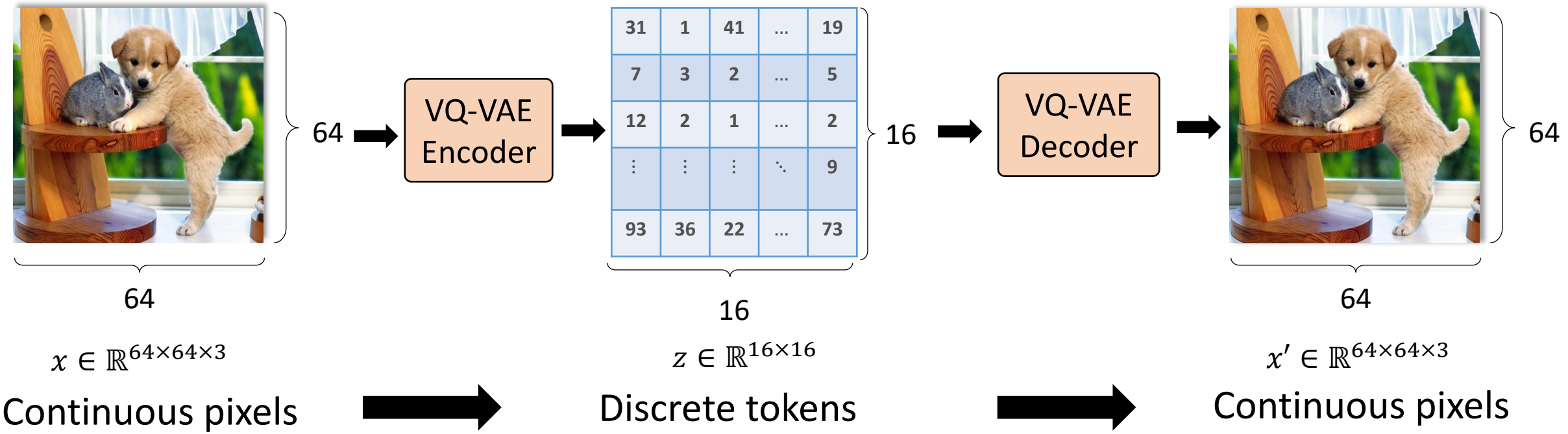
Discrete Visual Token



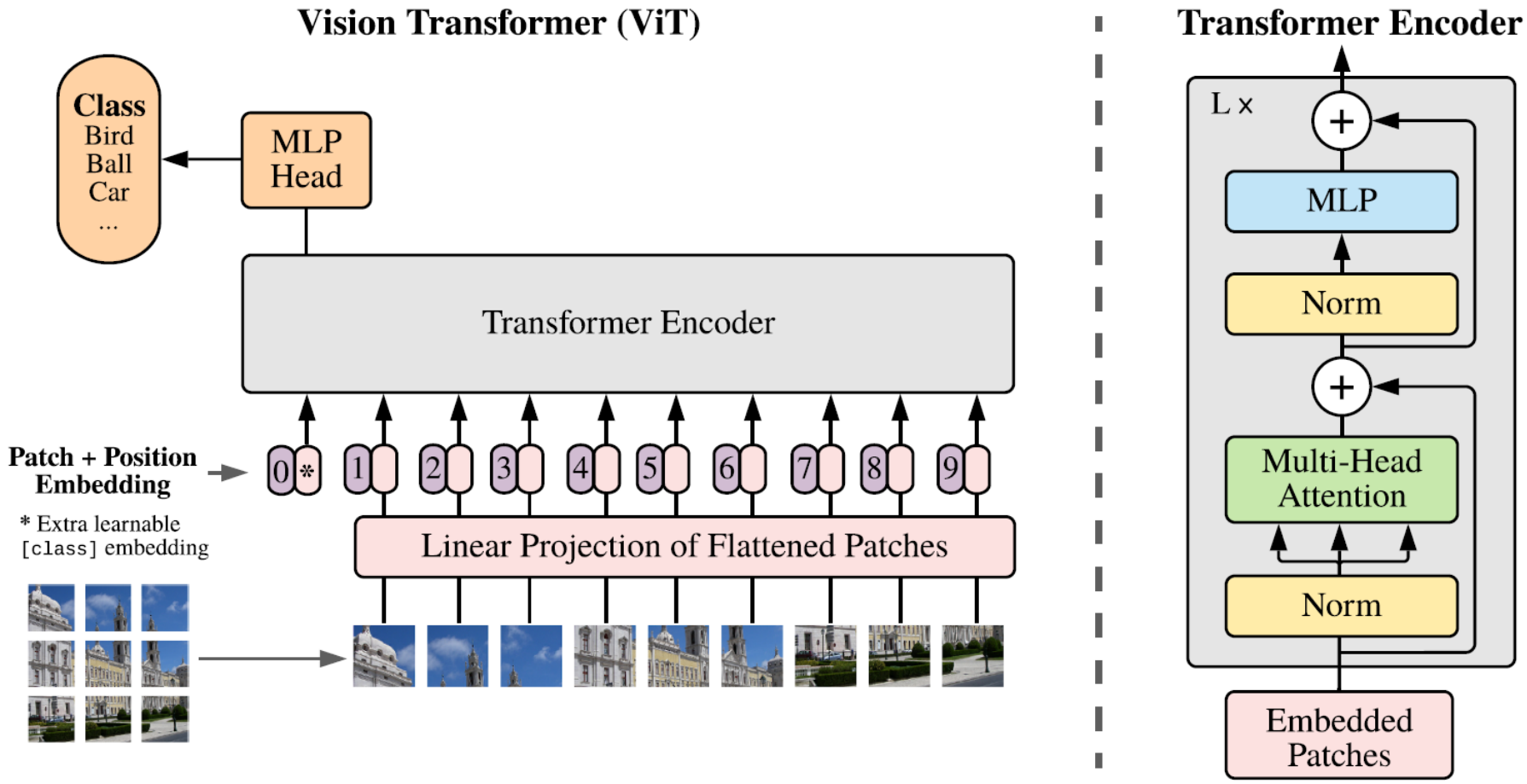
Object



VQ-VAE for Discrete Image Representation

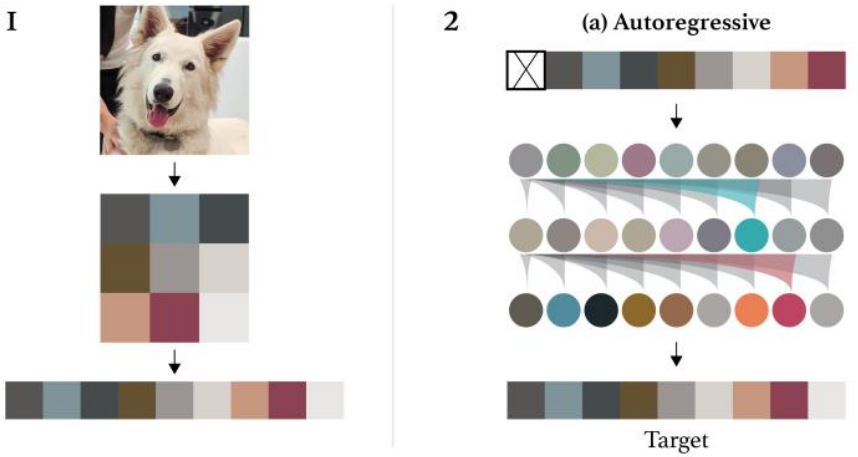


Vision Pre-training with Transformer and ImageNet

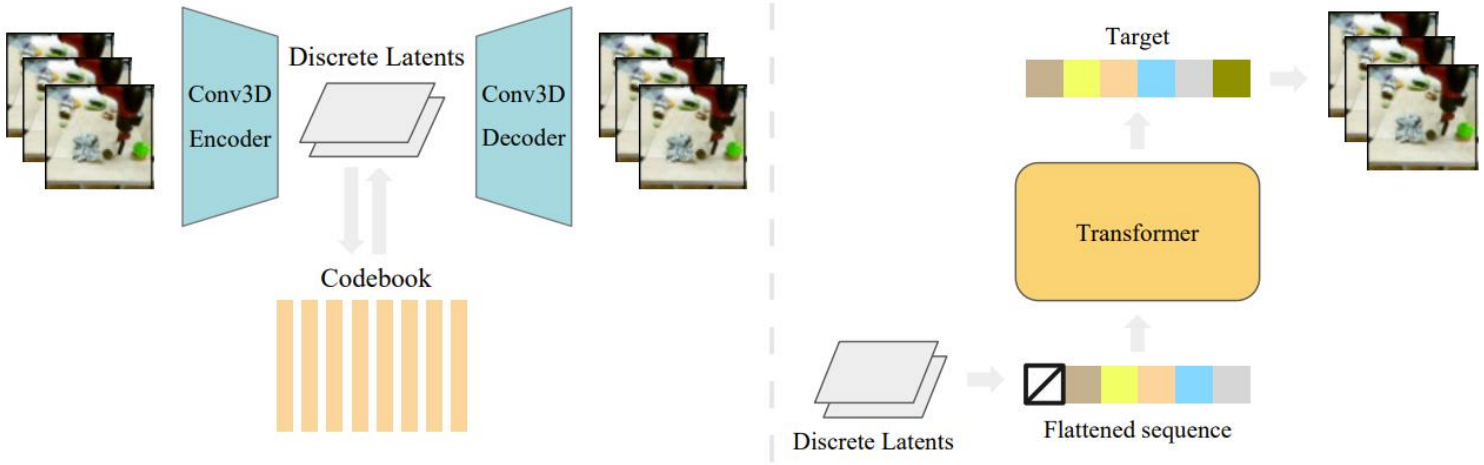


ViT (Dosovitskiy et al., 2021)
patch-based

Vision Pre-training with Auto-regressive Decoding

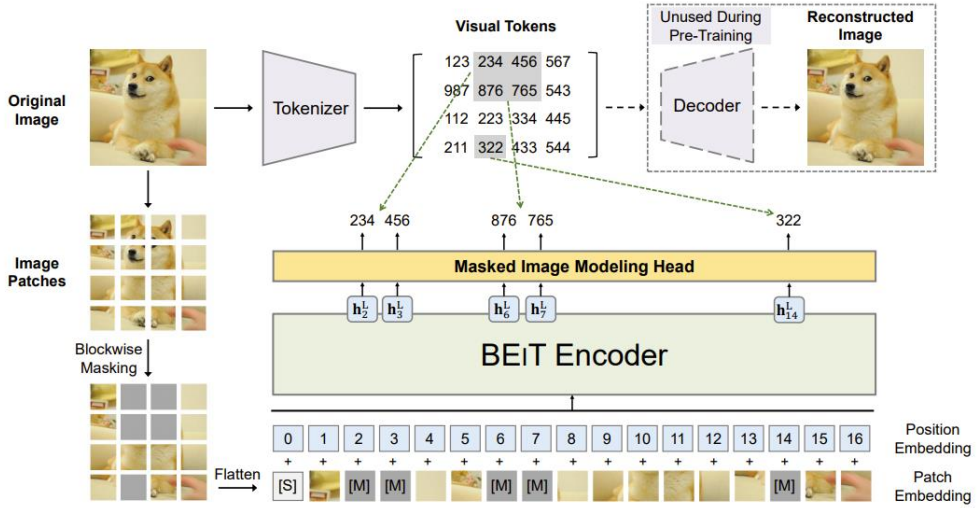


iGPT (Chen et al., 2020)
pixel-based

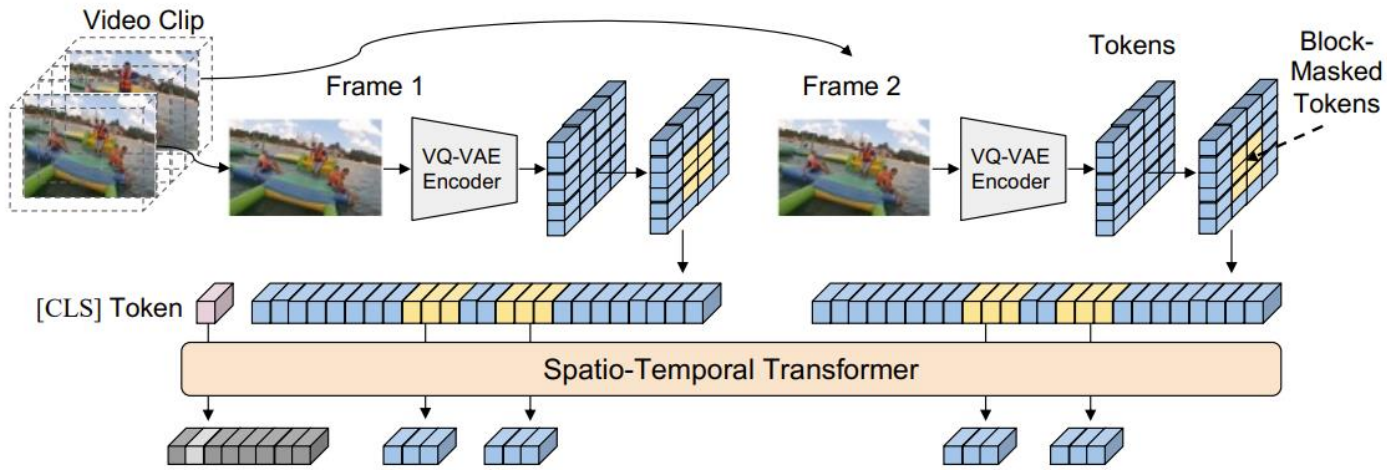


VideoGPT (Yan et al., 2021)
discrete visual token-based

Vision Pre-training with Denoising Auto-encoding

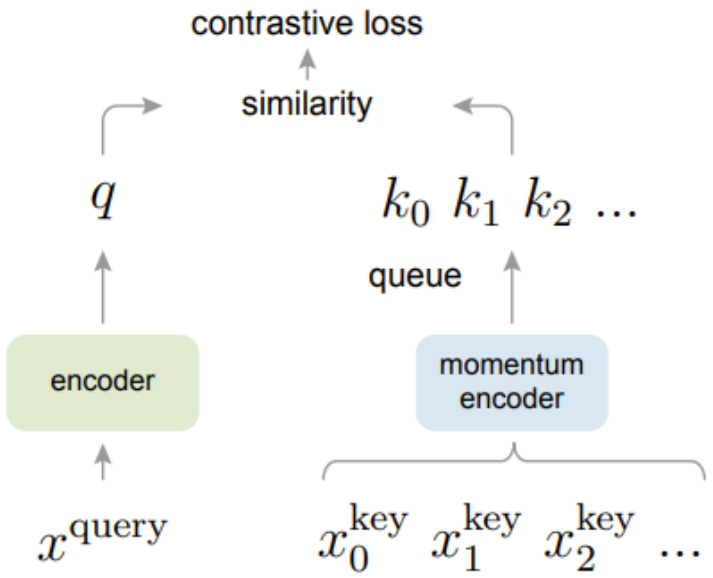


BEiT (Bao et al., 2021)
discrete visual token-based



VIMPAC (Tan et al., 2021)
discrete visual token-based

Vision Pre-training with Contrastive Learning



Algorithm 1 MoCo v3: PyTorch-like Pseudocode

```
# f_q: encoder: backbone + pred mlp + proj mlp
# f_k: momentum encoder: backbone + pred mlp
# m: momentum coefficient
# tau: temperature

for x in loader: # load a minibatch x with N samples
   x1, x2 = aug(x), aug(x) # augmentation
   q1, q2 = f_q(x1), f_q(x2) # queries: [N, C] each
   k1, k2 = f_k(x1), f_k(x2) # keys: [N, C] each

   loss = ctr(q1, k2) + ctr(q2, k1) # symmetrized
   loss.backward()

   update(f_q) # optimizer update: f_q
   f_k = m*f_k + (1-m)*f_q # momentum update: f_k

# contrastive loss
def ctr(q, k):
   logits = mm(q, k.t()) # [N, N] pairs
   labels = range(N) # positives are in diagonal
   loss = CrossEntropyLoss(logits/tau, labels)
   return 2 * tau * loss
```

Notes: `mm` is matrix multiplication. `k.t()` is `k`'s transpose. The projection head is excluded from `f_k` (and thus the momentum update).

SimCLR (Chen et al., 2020)

MoCo (He et al., 2020)

MoCo v3 (Chen et al., 2021)

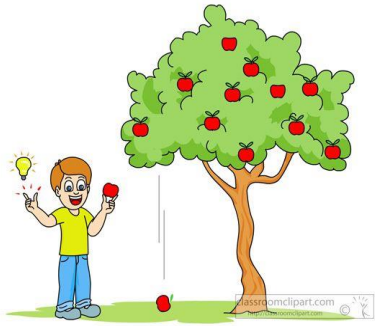
Outline

- Language Pre-training
- Vision Pre-training
- Vision-Language Pre-training
- Language-enhanced CV
- Vision-enhanced NLP
- Summarization

Multimodal AI is the new frontier.

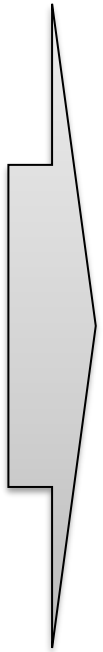
reading

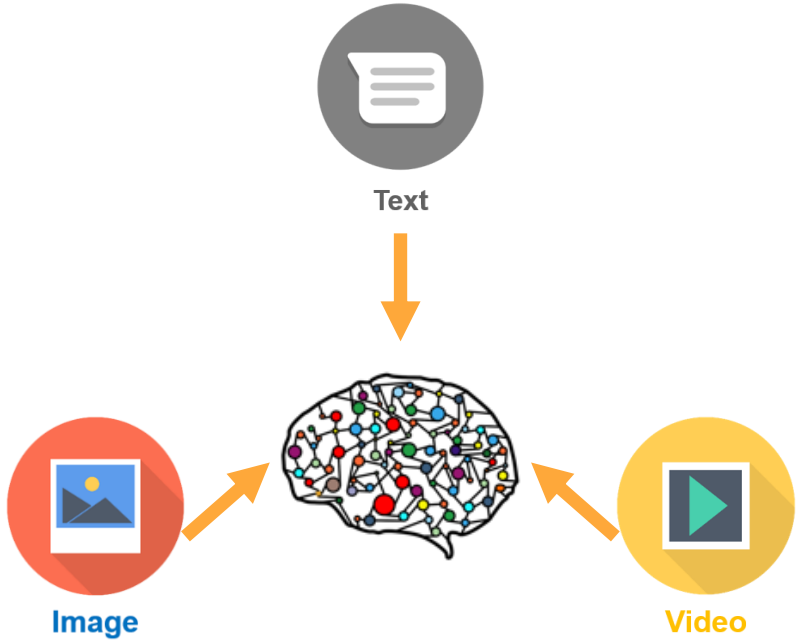
seeing



...

...

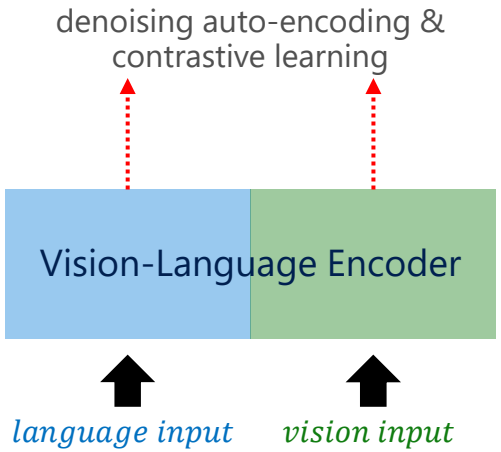




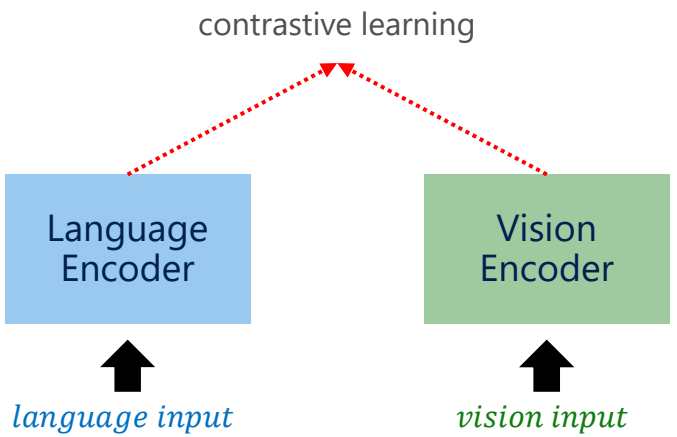
Humans learn from multiple senses.

Models learn from multimodal data.

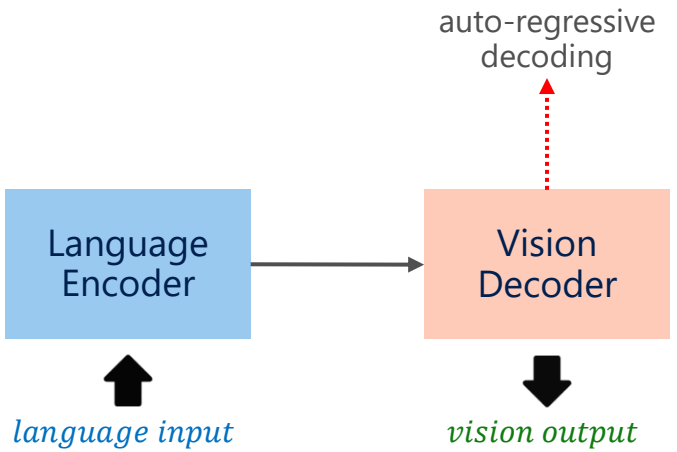
From Unimodal Pre-training to Vision-Language Pre-training



Single-stream

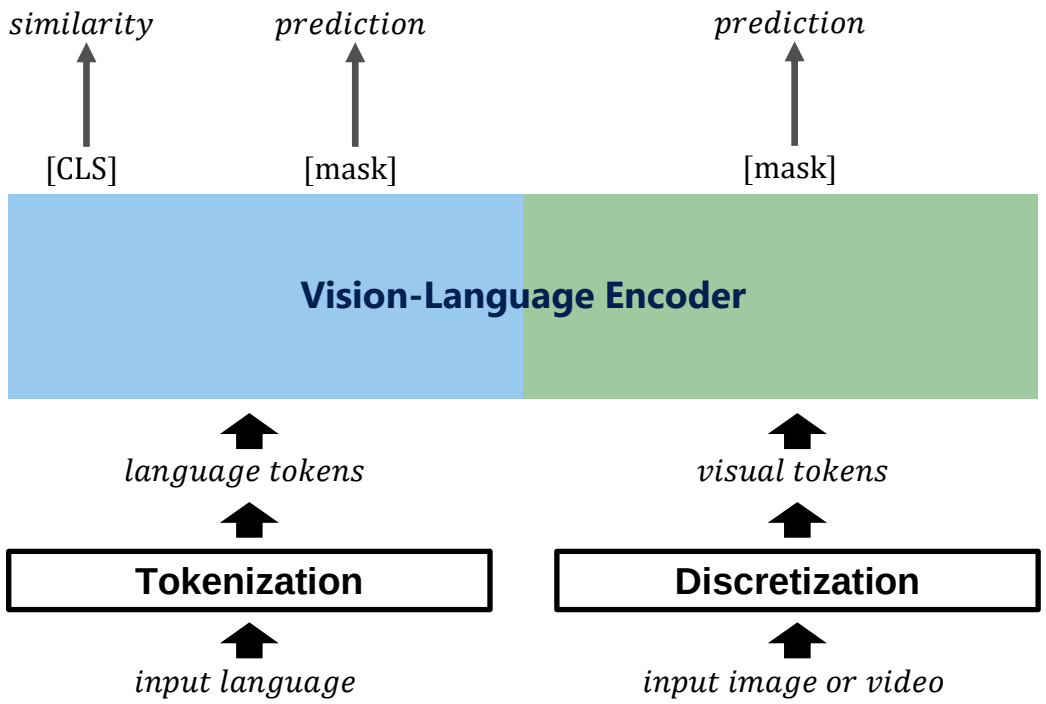


Two-stream



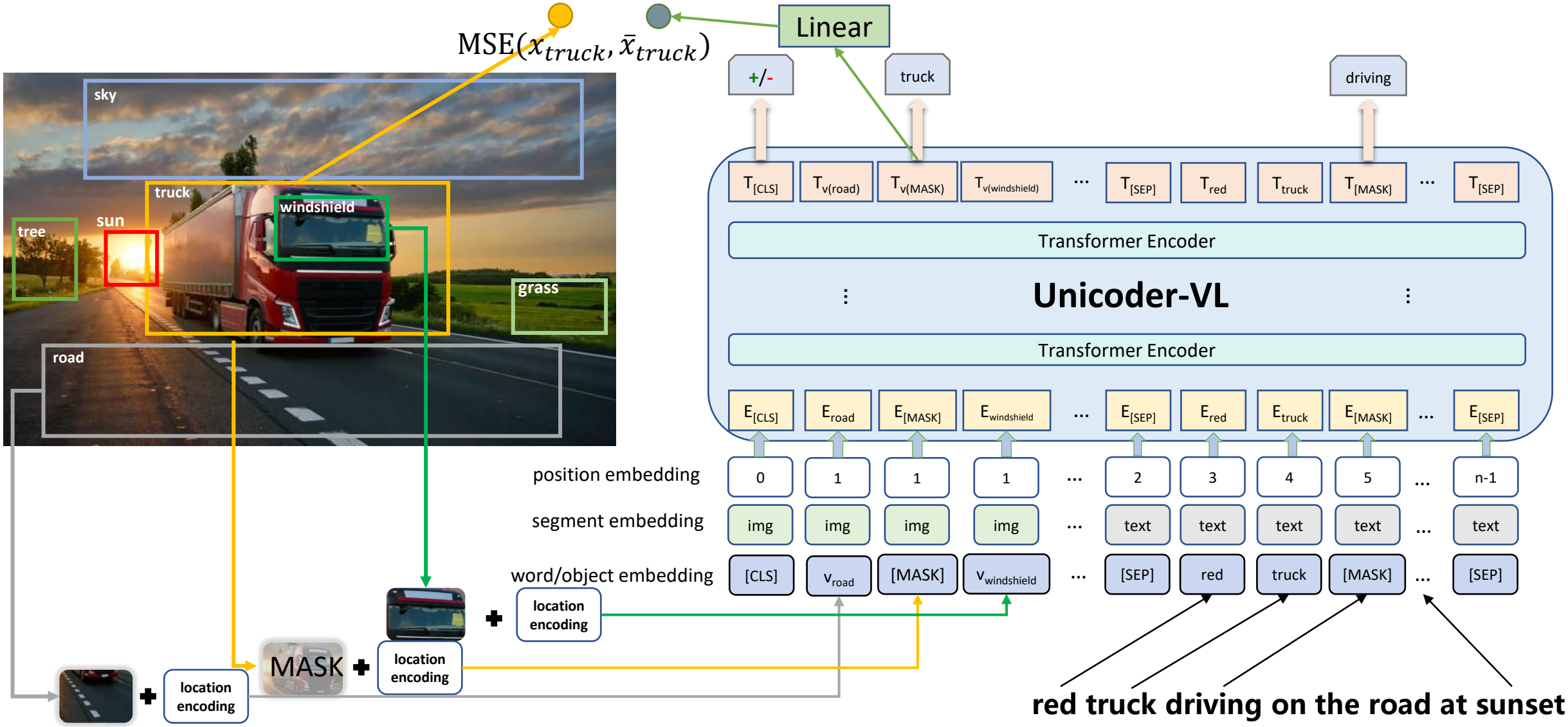
Encoder-Decoder

(1) One-stream Model



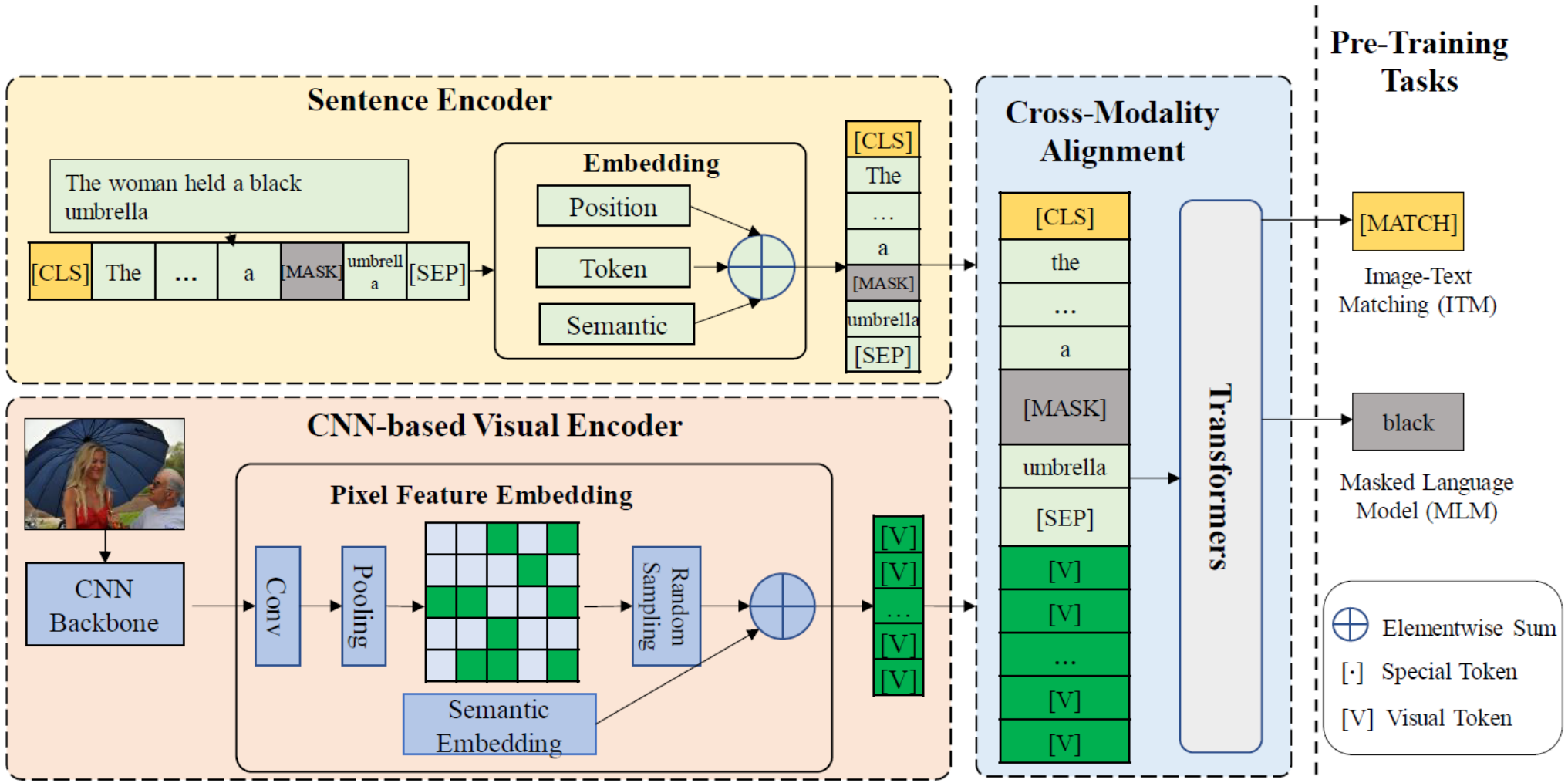
- **Vision representation**
  - Object (e.g., Unicoder-VL)
  - Pixel (e.g., Pixel-BERT)
  - Patch (e.g., ViLT)
  - Discrete visual token – VQ-VAE (e.g., SOHO)
  - Discrete visual token – Visual Clustering (e.g., VideoBERT)
- **Pre-training task**
  - Image-Text matching
  - Masked language modeling
  - Masked region modeling
  - Contrastive learning
- **Pre-training corpus**
  - Image-Text pairs + Multilingual sentences (e.g., M3P)
  - Without Image-Text pairs (e.g., U-VisualBERT)

Object-based Model: Unicoder-VL



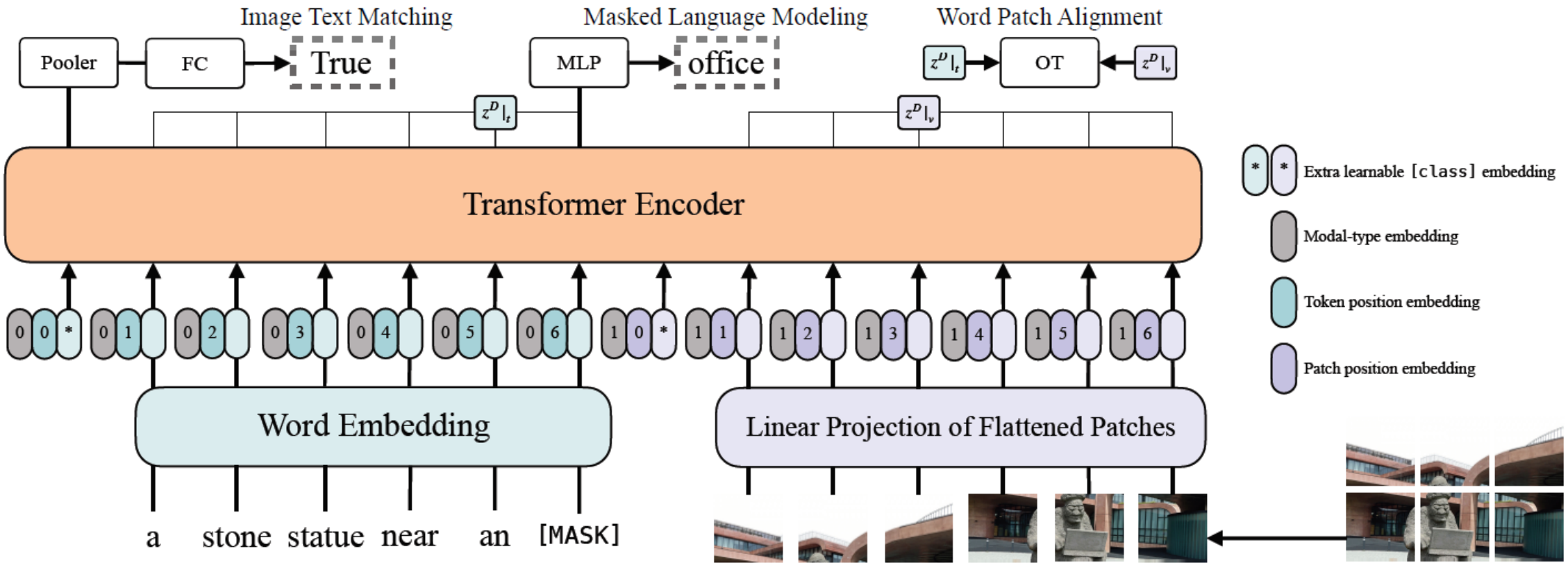
Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, Ming Zhou. **Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training**. AAAI 2020.

Pixel-based Model: Pixel-BERT



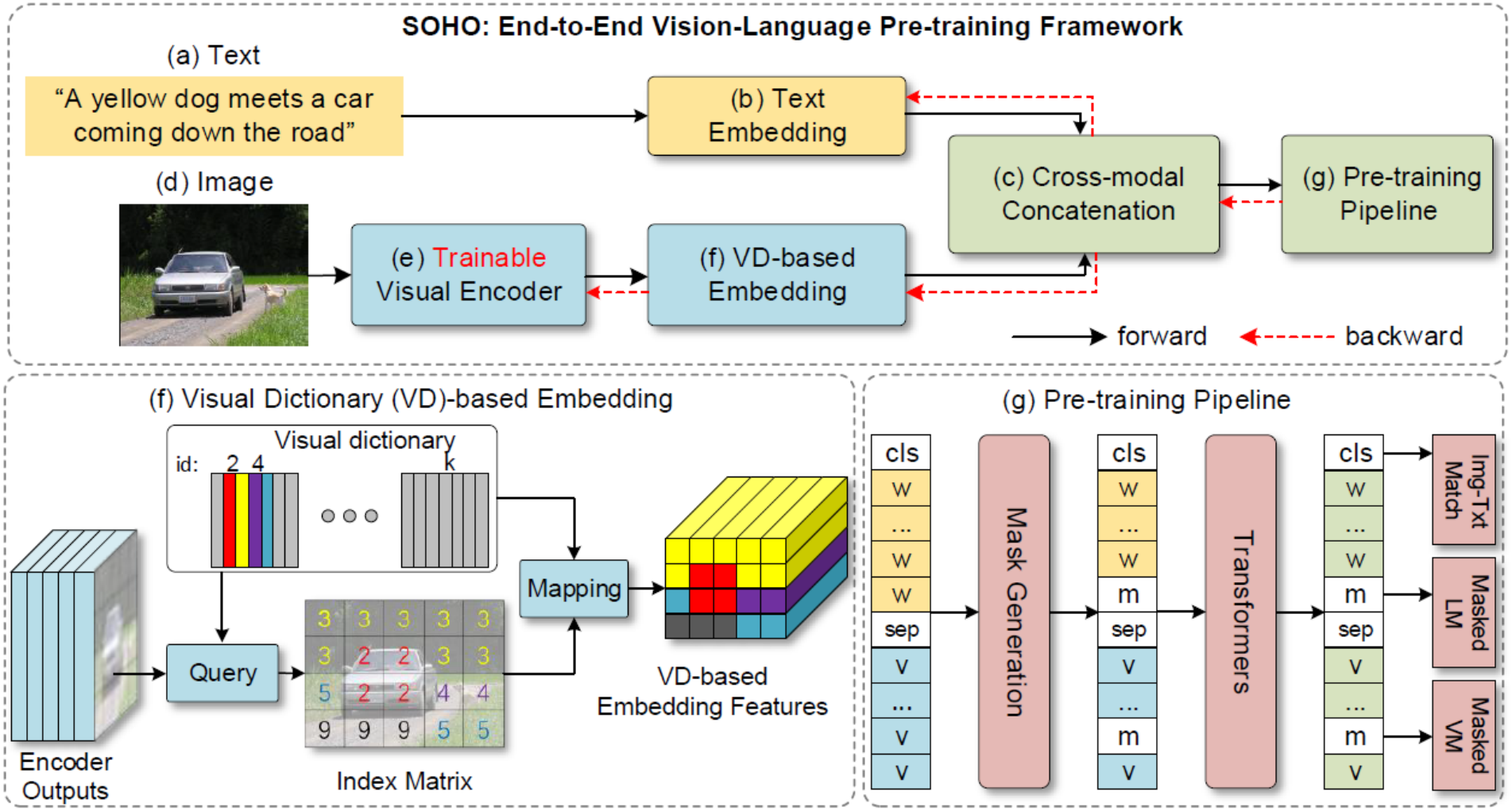
Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, Jianlong Fu. **Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers**. arXiv 2020.

Patch-based Model: ViLT



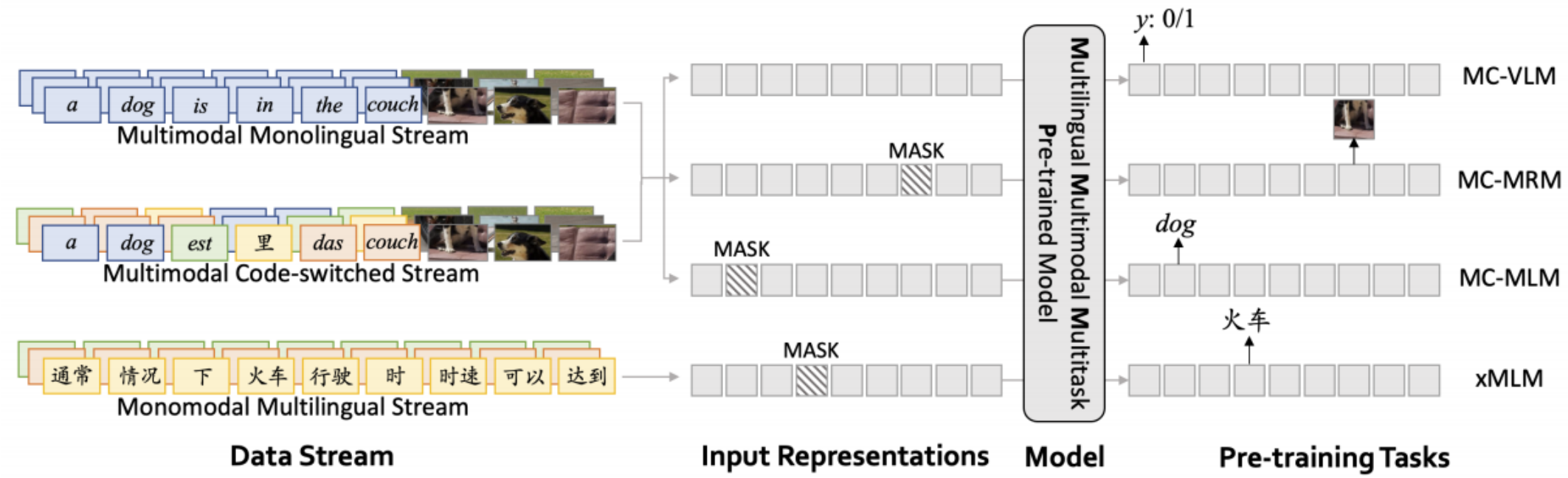
Wonjae Kim, Bokyung Son, Ildoo Kim. **ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision**. ICML 2021.

Discrete Visual Token-based Model: SOHO



Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, Jianlong Fu. **Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning**. CVPR 2021.

Model with Multilingual Corpus: M3P



Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, Nan Duan. **M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training**. CVPR 2021.

Evaluation Results

Task	Multilingual Image-Text Retrieval (Multi30K + MSCOCO)						Multilingual Image Captioning (Multi30K + MSCOCO)						Multimodal MT (Multi30K)	
	en	de	fr	cs	ja	zh	en	de	fr	cs	ja	zh	en→fr	en→de
SoTA	**92.7**	72.1	65.9	64.8	76.0	74.8	**37.4**	3.8	5.0	2.8	38.5	36.7	53.8	31.6
M^3P_B	88.0	**82.0**	**73.5**	**70.2**	**86.8**	**81.8**	34.7	**16.6**	**8.7**	**5.4**	**40.2**	**39.7**	**55.5**	**35.7**
Δ	4.7 ↓	9.9 ↑	7.6 ↑	5.4 ↑	10.8 ↑	7.0 ↑	3.7 ↓	12.8 ↑	3.7 ↑	2.6 ↑	1.7 ↑	3.0 ↑	1.7 ↑	4.1 ↑

Blue numbers indicates the best result for a task. For retrieval tasks, we use *mean Recall* as the metric, which is an average score of R@1, R@5 and R@10 on i2t and t2i tasks. For captioning and translation tasks, we use *BLEU-4* as the metric.

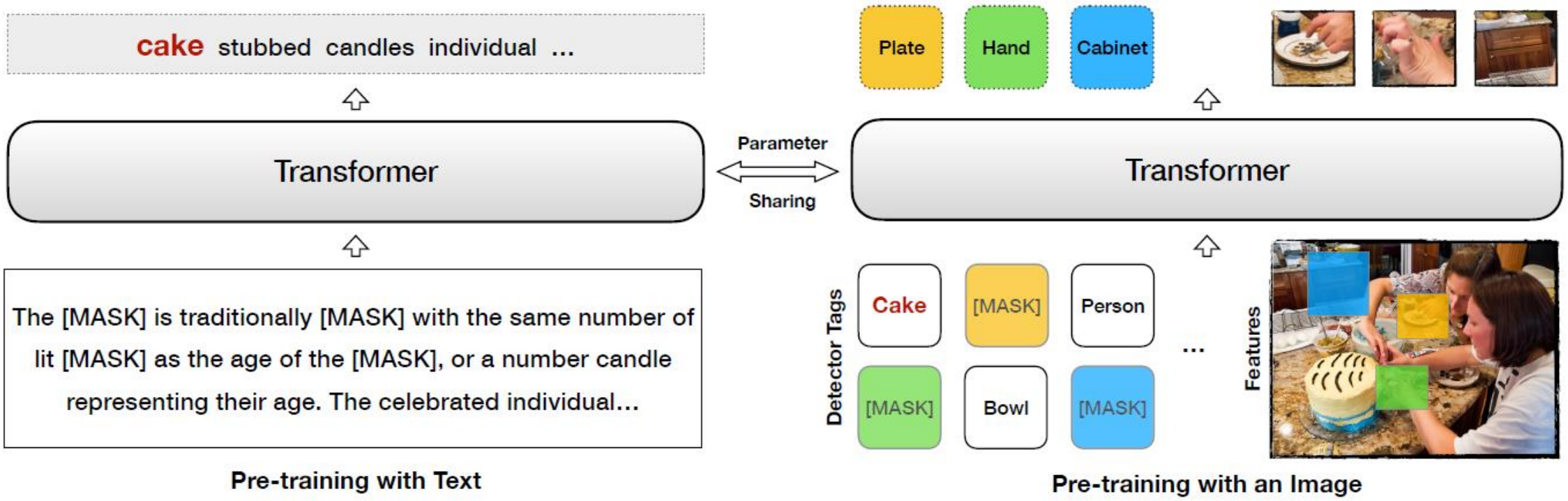
image caption output (**zh**): 一辆载着人和纸糊的房子的卡车行驶在街道上
(translation: a truck carrying people and paper houses travels down the street)

image caption input (**en**): A Boston Terrier is running on lush green grass in front of a white fence.

caption translation output (**fr**): Le Boston Terrier court sur l'herbe verte luxurie devant une clôture blanche.
(translation: The Boston Terrier runs on lush green grass in front of a white fence.)

caption translation output (**de**): Ein Hund läuft auf grünem Rasen vor einem weißen Zaun.
(translation: A dog runs on green grass in front of a white fence.)

Model without Image-Text Pairs: U-VisualBERT



Liunian Harold Liy, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, Kai-Wei Chang. **Unsupervised Vision-and-Language Pre-training Without Parallel Images and Captions**. NAACL 2021.

Evaluation Results

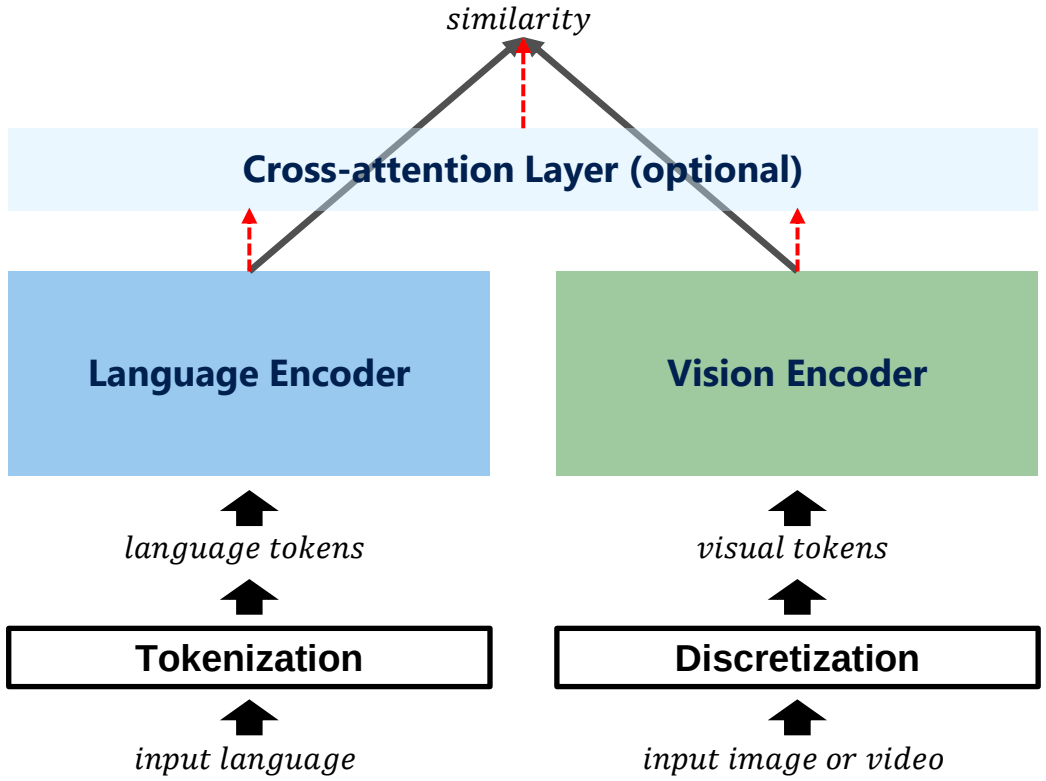
Model	Aligned	Unaligned		VQA	NLVR^2		Flickr30K			RefCOCO+		
		Image	Text	Test-Dev	Dev	Test-P	R@1	R@5	R@10	Dev	TestA	TestB
Pre-BERT	-	-	-	70.22	54.1	54.8	48.60	77.70	85.20	65.33	71.62	56.02
ViLBERT	3M	0	0	70.55	-	-	58.78	85.60	91.42	72.34	78.52	62.61
VL-BERT	3M	0	~50M	71.16	-	-	-	-	-	71.60	77.72	60.99
UNITER$_{cc}$	3M	0	0	**71.22**	-	-	-	-	-	72.49	79.36	63.65
S-VisualBERT	3M	0	2.5M	70.87$_{\pm.02}$	**73.44**$_{\pm.51}$	**73.93**$_{\pm.51}$	**61.19**$_{\pm.06}$	**86.32**$_{\pm.12}$	**91.90**$_{\pm.02}$	**73.65**$_{\pm.11}$	**79.48**$_{\pm.36}$	**64.49**$_{\pm.22}$
Base	0	0	0	69.26	68.40	68.65	42.86	73.62	83.28	70.66	77.06	61.43
U-VisualBERT	0	3M	5.5M	**70.74**$_{\pm.06}$	**71.74**$_{\pm.24}$	**71.02**$_{\pm.47}$	**55.37**$_{\pm.49}$	**82.93**$_{\pm.07}$	**89.84**$_{\pm.21}$	**72.42**$_{\pm.06}$	**79.11**$_{\pm.08}$	**64.19**$_{\pm.54}$

Table 1: Evaluation results on four V&L benchmarks. Our unsupervised model trained with unaligned data (U-VisualBERT) achieves close performance with a supervised model trained with aligned data (S-VisualBERT). U-VisualBERT also rivals with several supervised models such as ViLBERT on most metrics.

Model	Text		VQA	NLVR^2		Flickr30K			RefCOCO+		
	Caption	General	Test-Dev	Dev	Test-P	R@1	R@5	R@10	Dev	TestA	TestB
Base	-	-	69.26	68.40	68.65	42.86	73.62	83.28	70.66	77.06	61.43
U-VisualBERT	CC	BC	**70.74**	71.74	71.02	55.37	**82.93**	89.84	72.42	79.11	64.19
U-VisualBERT$_{SBU}$	SBU	BC	70.70	**71.97**	**72.11**	**56.12**	82.82	**90.12**	**73.05**	**79.48**	64.19
U-VisualBERT$_{NC}$	-	BC	70.47	71.47	71.19	54.36	82.22	89.24	72.96	79.30	**64.25**

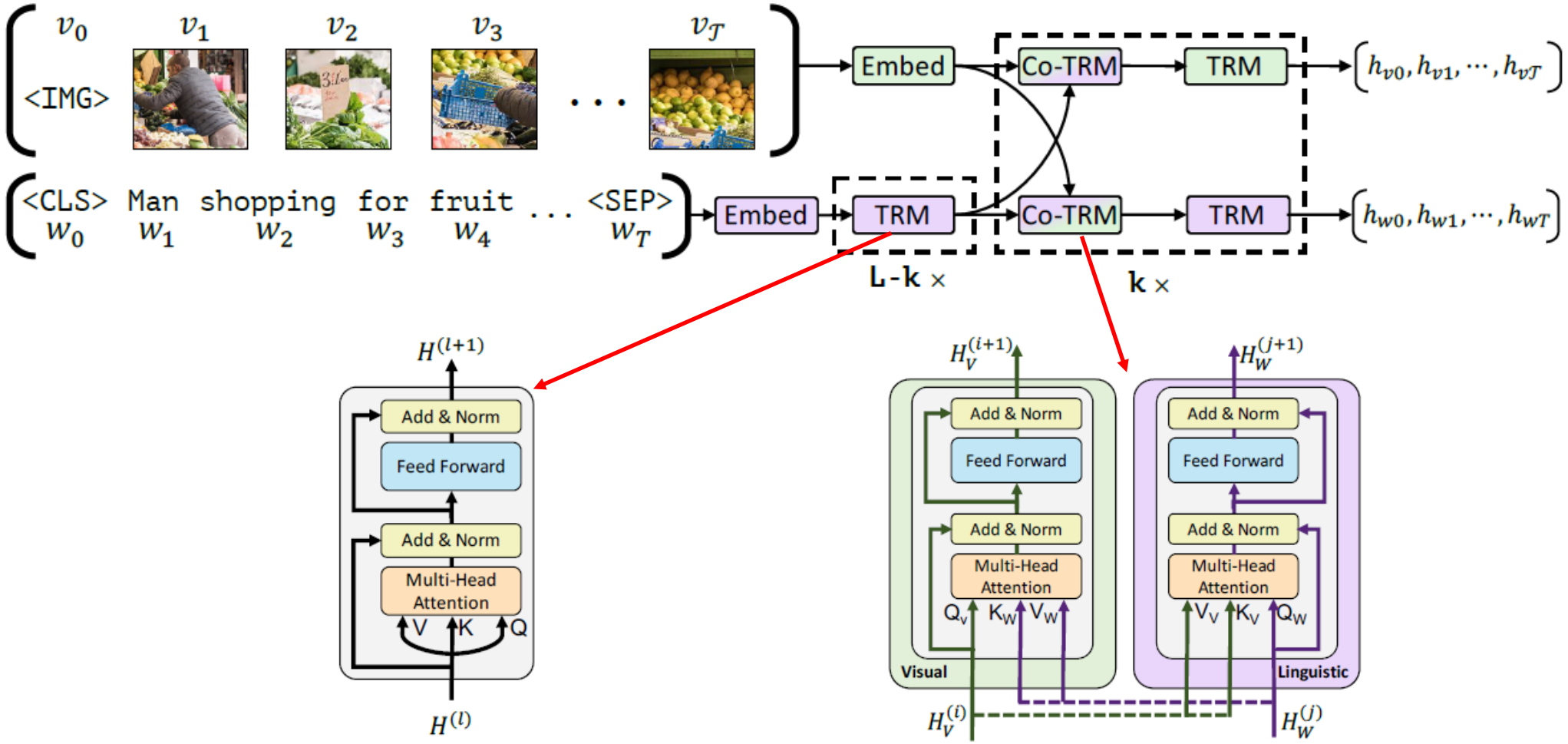
Table 2: Unsupervised pre-training is applicable when images and captions are collected independently (U-VisualBERT$_{SBU}$) or when no caption text is provided (U-VisualBERT$_{NC}$).

(2) Two-stream Model



- **Image-Text pre-training**
  - ViLBERT
  - CLIP
- **Video-Text pre-training**
  - UniVL
  - CLIP4Clip

Image-Text Pre-training: ViLBERT

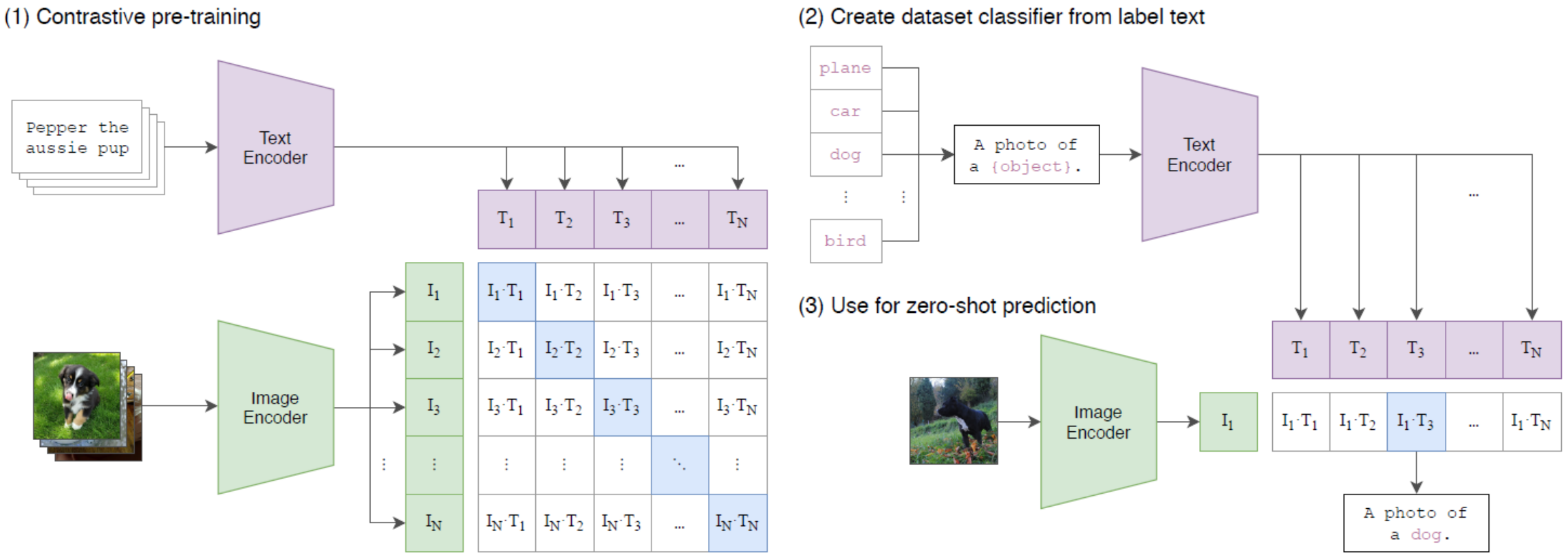


(a) Standard encoder transformer block

(b) Our co-attention transformer layer

Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee. **ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks**. NeurIPS 2019.

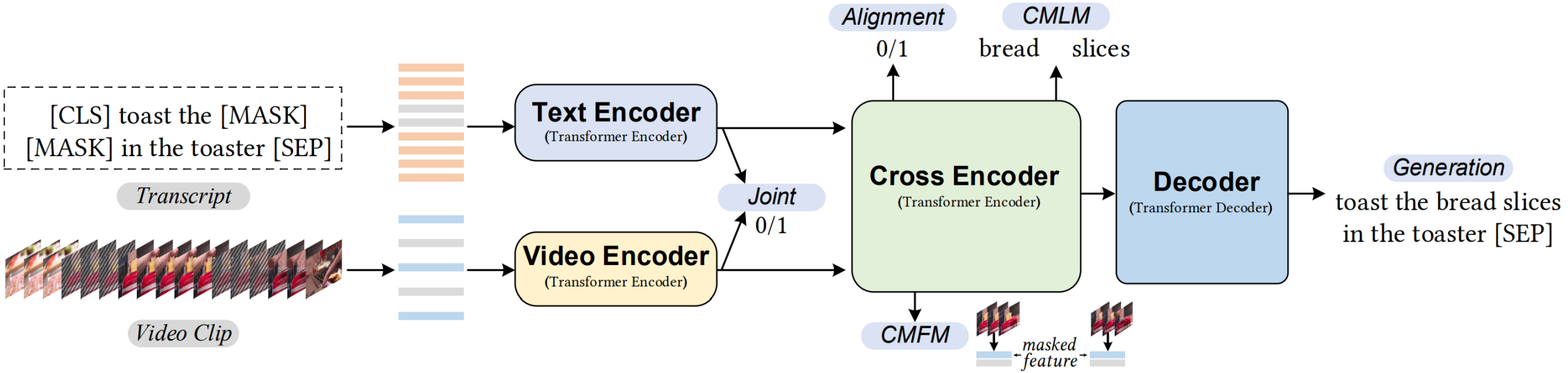
Image-Text Pre-training: CLIP

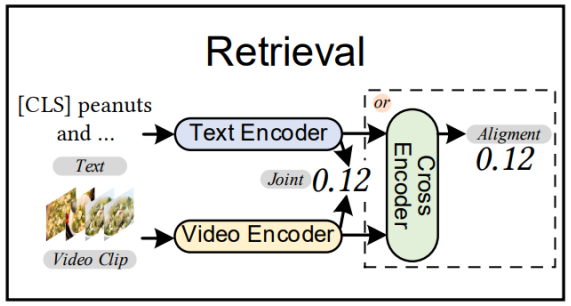


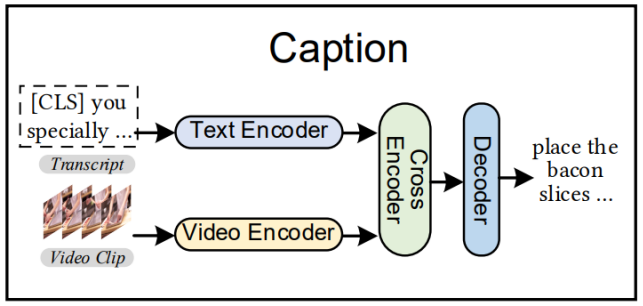
Alec Radford et al. **Learning Transferable Visual Models From Natural Language Supervision**. ICML 2021.

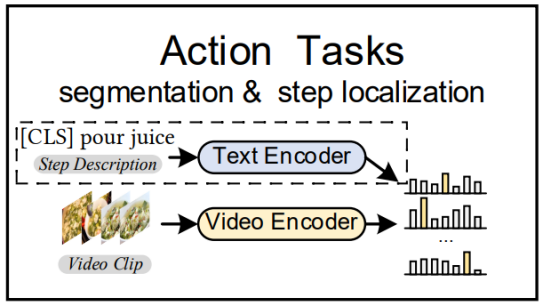
Video-Text Pre-training: UniVL

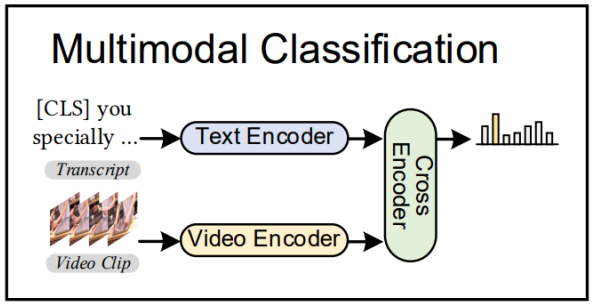
1. Video-Text Joint Embedding **2.** Video-Text Alignment **3.** Masked Frame Model **4.** Masked Language Model **5.** Caption Generation





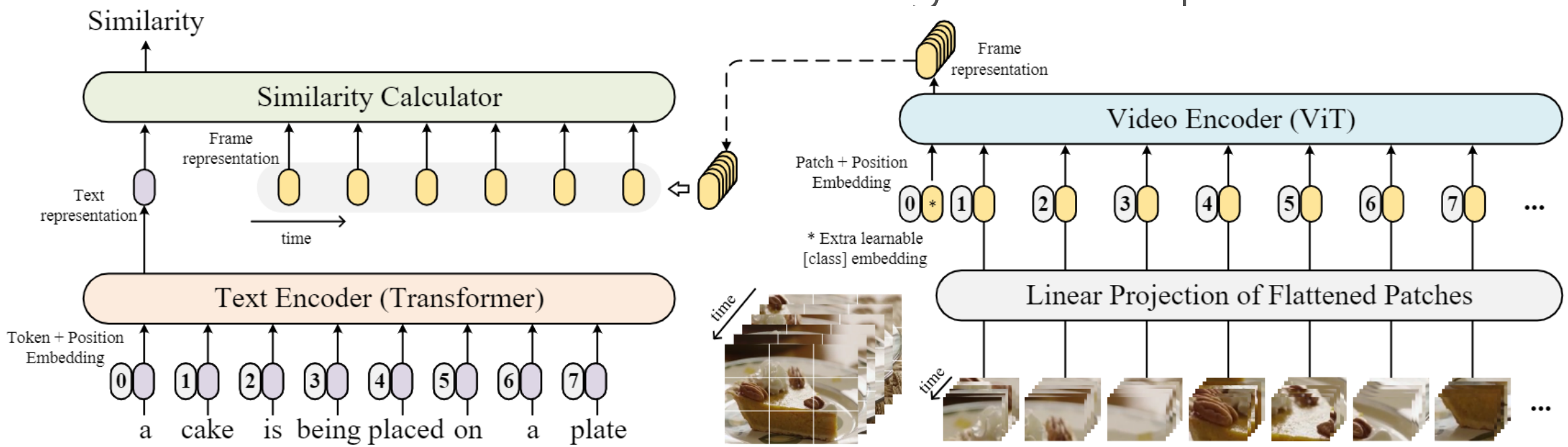


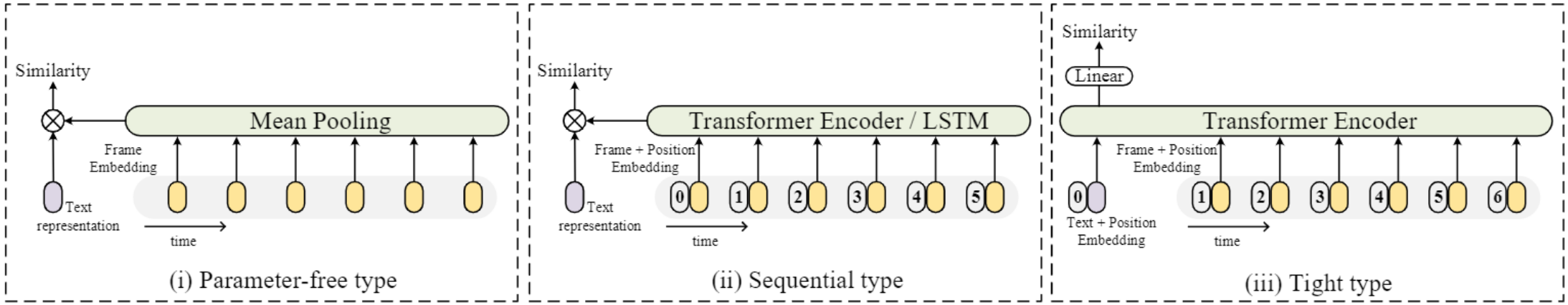




Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, Ming Zhou. **UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation**. arXiv 2020.

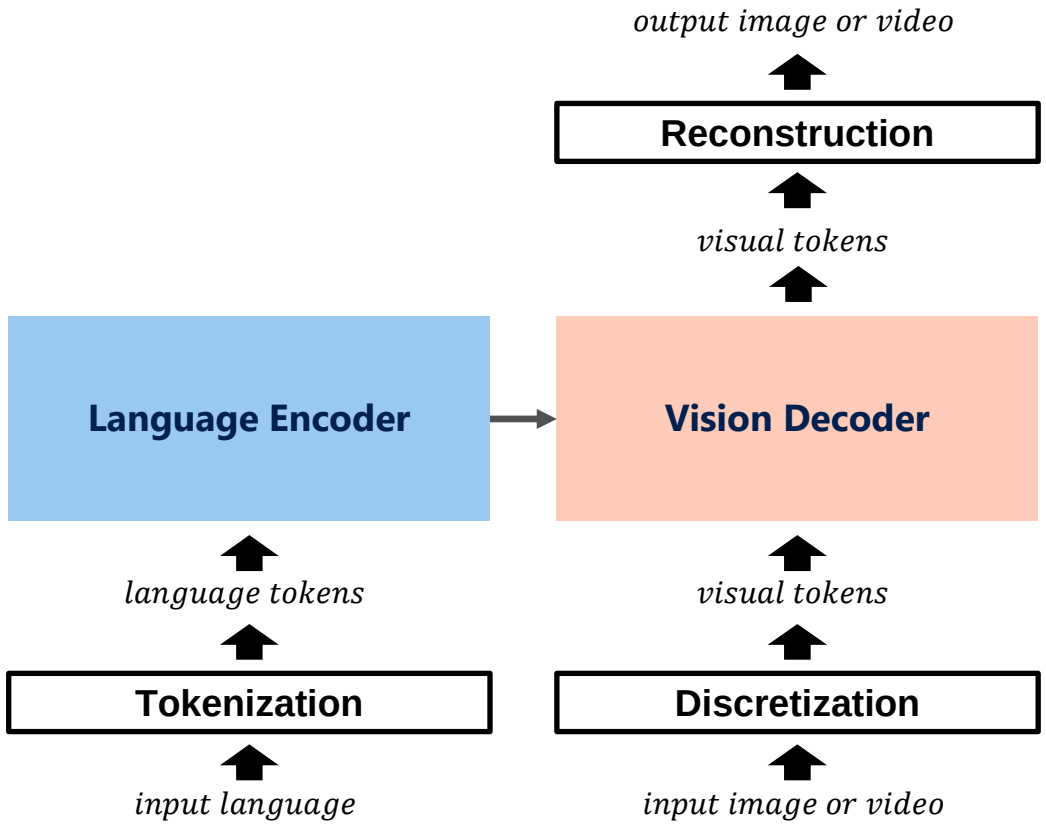
Video-Text Pre-training: CLIP4Clip





Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, Tianrui Li. **CLIP4Clip: An Empirical Study of CLIP for End-to-End Video Clip Retrieval**. arXiv 2021.

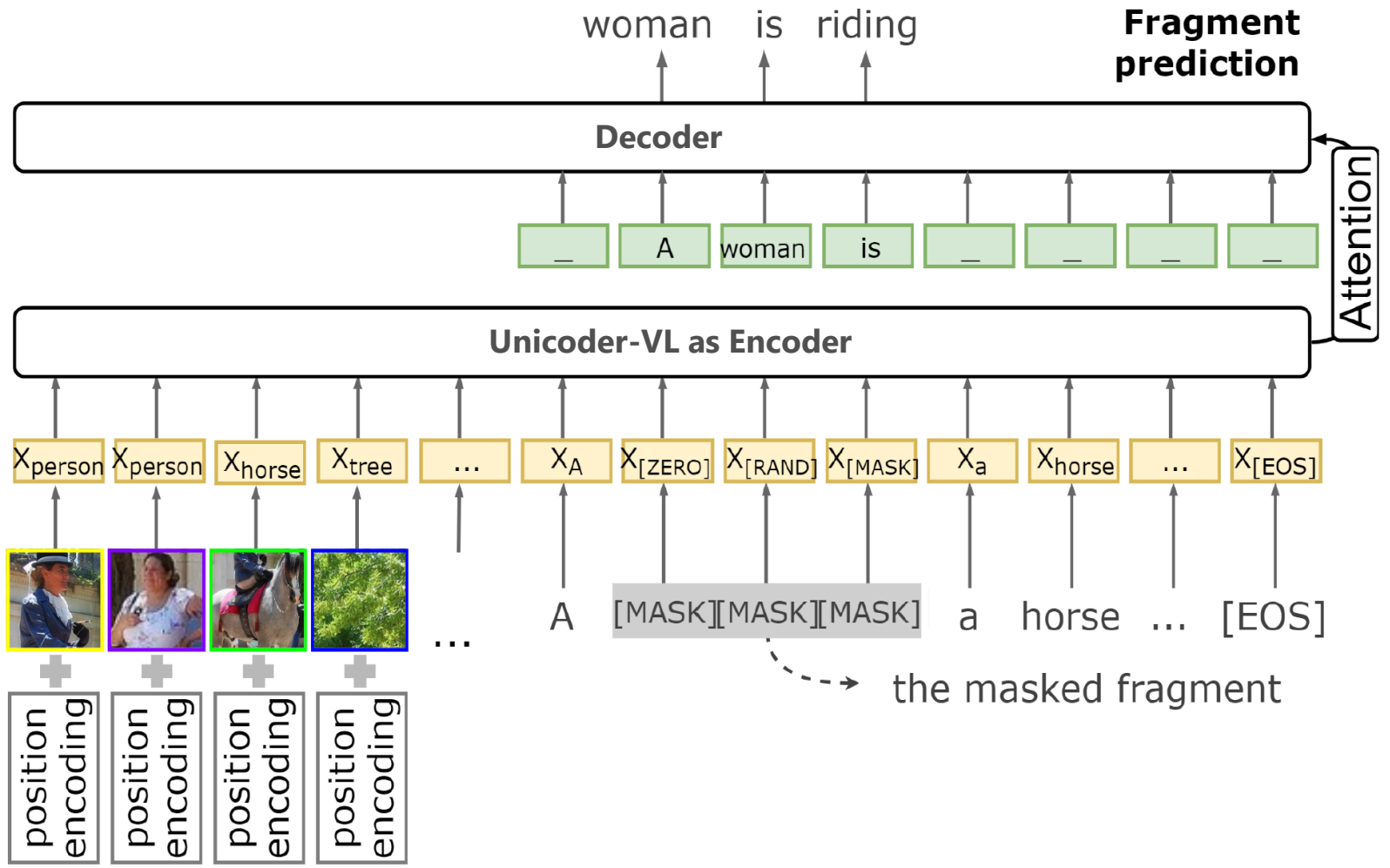
(3) Encoder-Decoder Model



- **Cross-modal generation**
  - Image-to-Text (e.g., XGPT, LoopCAG)
  - Video-to-Text (e.g., UniVL)
  - Text-to-Image (e.g., DALL·E)
  - Text-to-Video (e.g., GODIVA)

- **Open challenge**
  - Long sequence
  - High resolution
  - Spatial-Temporal modeling
  - Joint training with VQ-VAE/VQ-GAN
  - Limited vision-language pairs
  - Evaluation metric for generation tasks

Image-to-Text: XGPT



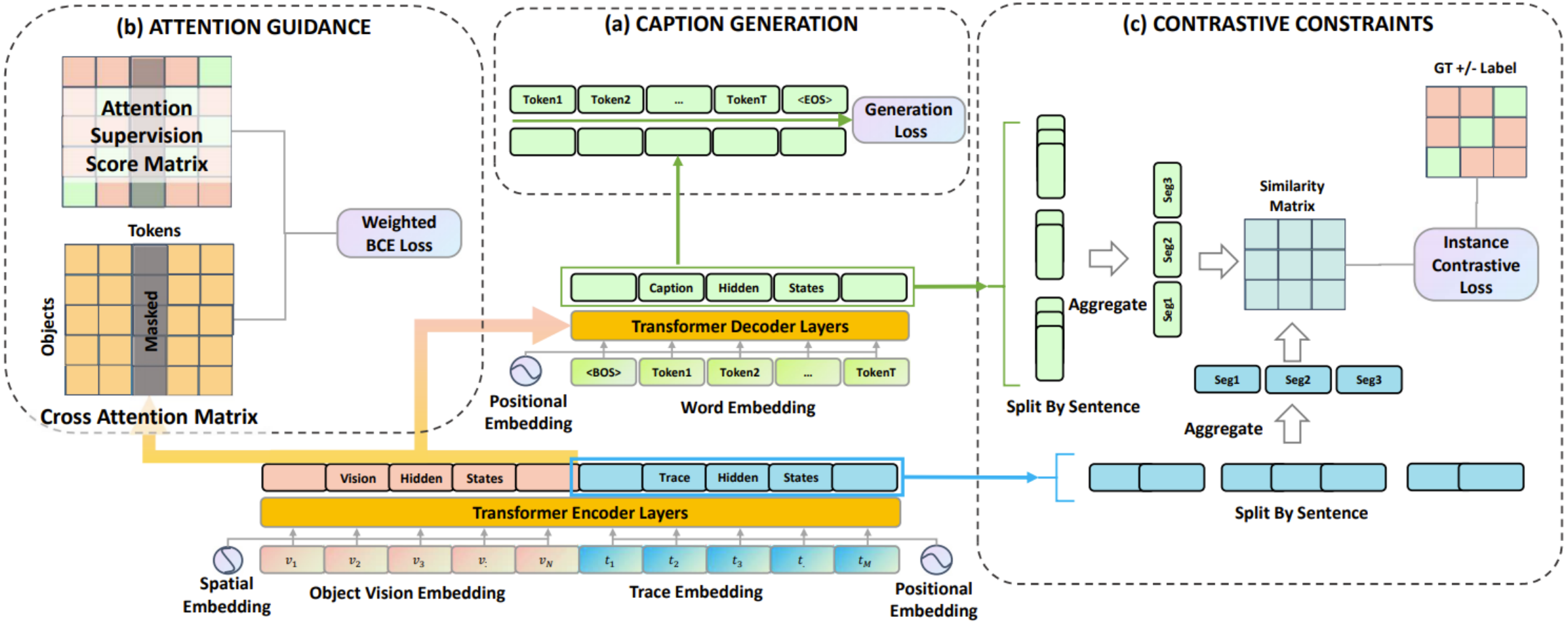
Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, Ming Zhou. **XGPT: Cross-modal Generative Pre-Training for Image Captioning**. NLPCC 2021.

Image-to-Text based on Trace

In this picture there is a stand on a ground. On the backside there is a person. He is riding on a horse. He is wearing a cap. He is in between the fence. There is a flags on a wall. On the left side there is a score board on a table and flower plants. We can see in the background sky and trees.

Jordi Pont-Tuset, Jasper Uijlings, Beer Changpinyo, Radu Soricut, Vittorio Ferrari. **Connecting vision and language with localized narratives**. ECCV 2020.

Image-to-Text: LoopCAG



Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, Shuai Ma. **Control Image Captioning Spatially and Temporally**. ACL 2021.

Text-to-Image: DALL·E



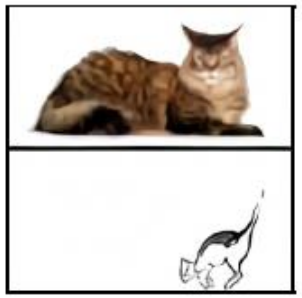

(a) a tapir made of accordion. a tapir with the texture of an accordion.




(b) an illustration of a baby hedgehog in a christmas sweater walking a dog



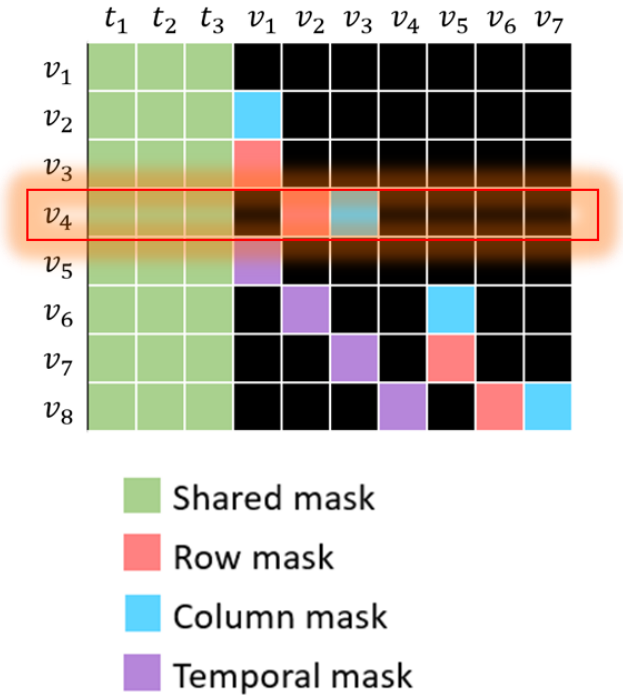
(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

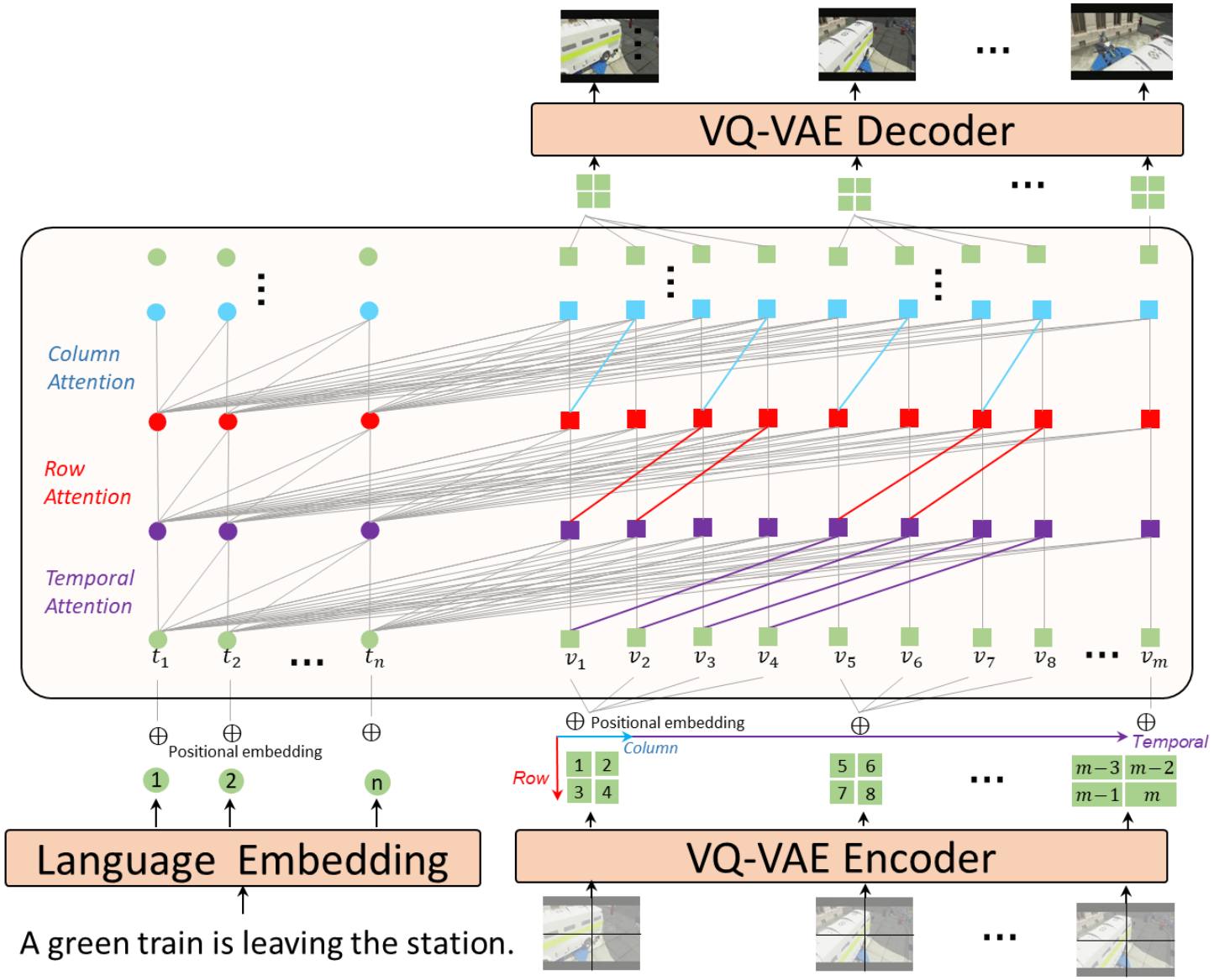
(d) the exact same cat on the top as a sketch on the bottom

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever. **Zero-Shot Text-to-Image Generation**. arXiv 2021.

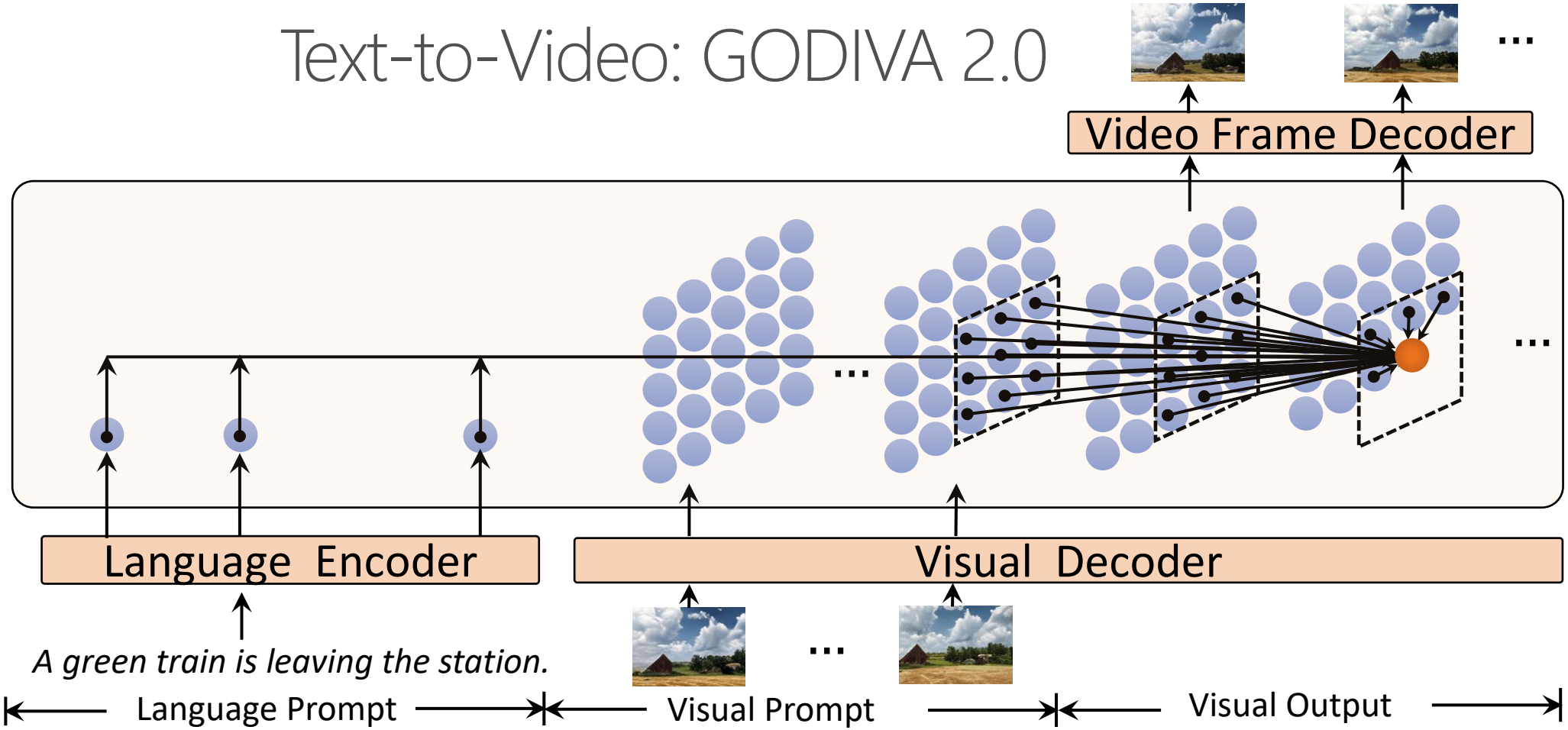
Text-to-Video: GODIVA 1.0



To generate v_4, our model attends to t_1, t_2, t_3 and v_2, v_3.



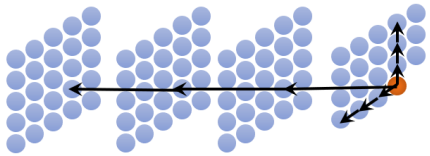
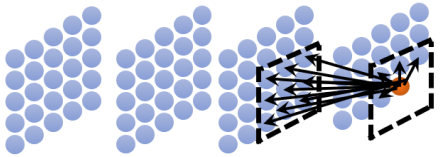
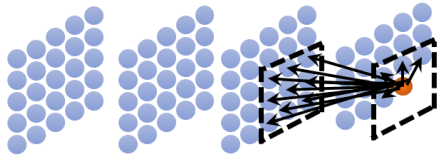
Text-to-Video: GODIVA 2.0



Multi-task Supporting

Input Text	Output Text	
Input Text	Output Image	
Input Text	Output Video	
Input Text	Input Image	Output Video
Input Text	Input Video	Output Video
Input Image	Output Video	
Input Video	Output Video	

Model Variations & Pre-training Settings

Main Differences	GODIVA_BASE (1.0)	GODIVA_MEDIUM (2.0)	GODIVA_LARGE (2.0)
Dataset	MSRVTT: 70K text-video pairs	MSRVTT: 70K text-video pairs VATEX: 20K text-video pairs COCO: 591K captions Moments: 727K videos	MSRVTT: 70K text-video pairs VATEX: 20K text-video pairs COCO: 591K text-image pairs Moments: 727K videos
Parameters	265M	709M	1.62B
Frame Resolution	64×64	256×256	256×256
Visual Backbone	VQ-VAE	VQ-GAN	VQ-GAN
Training Speed (64 A100 GPUs)	13 min/epoch, ~500 epochs to converge(4.5 days)	20 min/epoch, ~1000 epochs to converge (13 days)	40 min/epoch, nearly 1000 epochs to converge (27 days)
Key Architecture	3-D Axial Sparse Attention	3-D Nearby Sparse Attention	3-D Nearby Sparse Attention
			

Text as Prompt

Input Text

a person is preparing some art

Output Video

Input Text

a man with suit sitting on the chair talking in front of the camera

Output Video

a gamer talks about his Minecraft experience

a girl and someone is putting a painted egg in to a water

Fine-tune on Moving MNIST

Input Text

digit 9 is moving down then up

Output Video

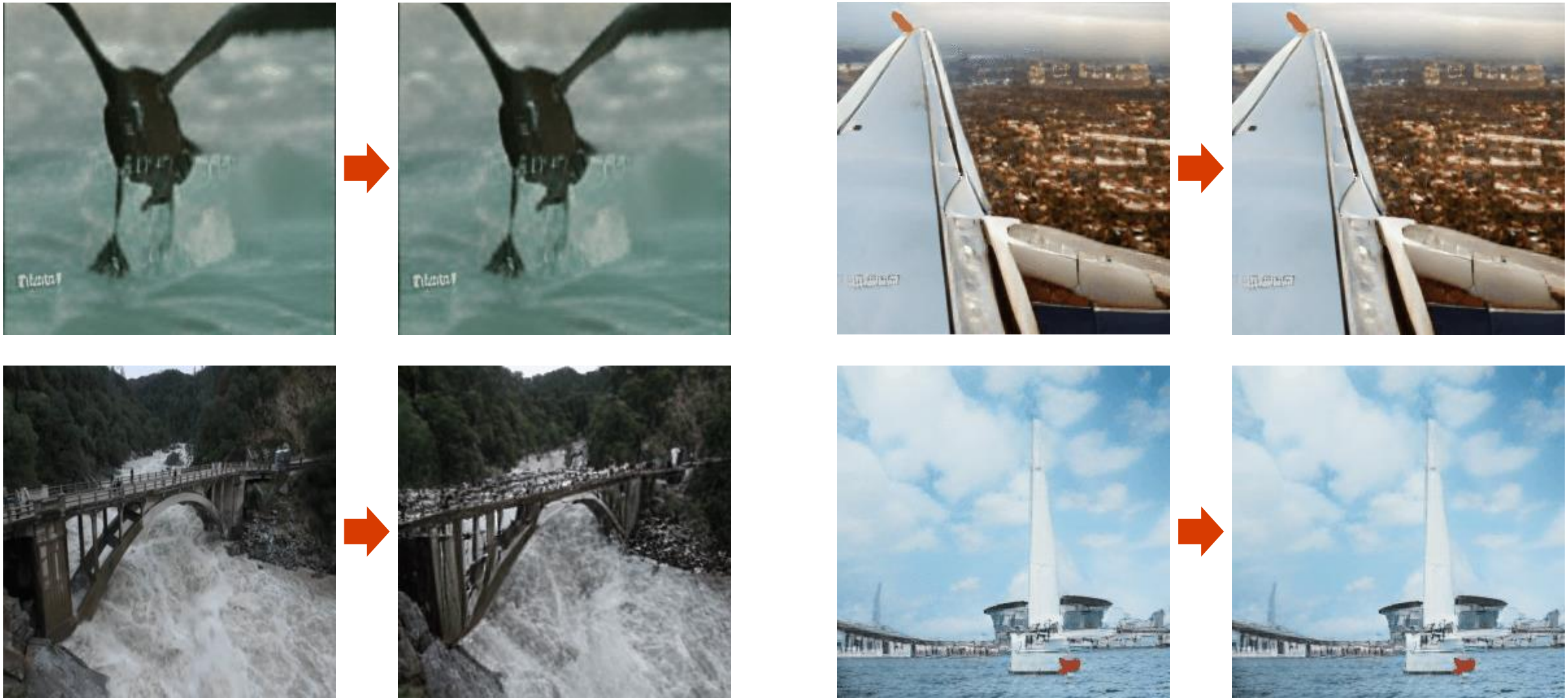
Input Text

digit 7 moves right then left while digit 3 moves down then up

Output Video

Image as Prompt

Input Image **Output Video** **Input Image** **Output Video**



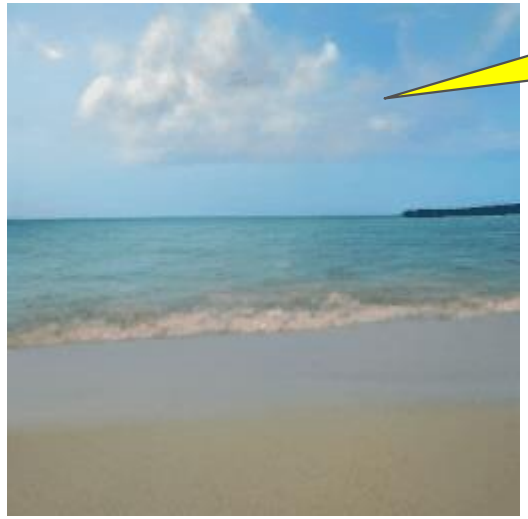
Zero-shot Examples

Input Image **Output Video** **Input Image** **Output Video**

Text + Image as Prompt

Input Text | **Input Image** | **Output Video**

there are clouds floating over the sea



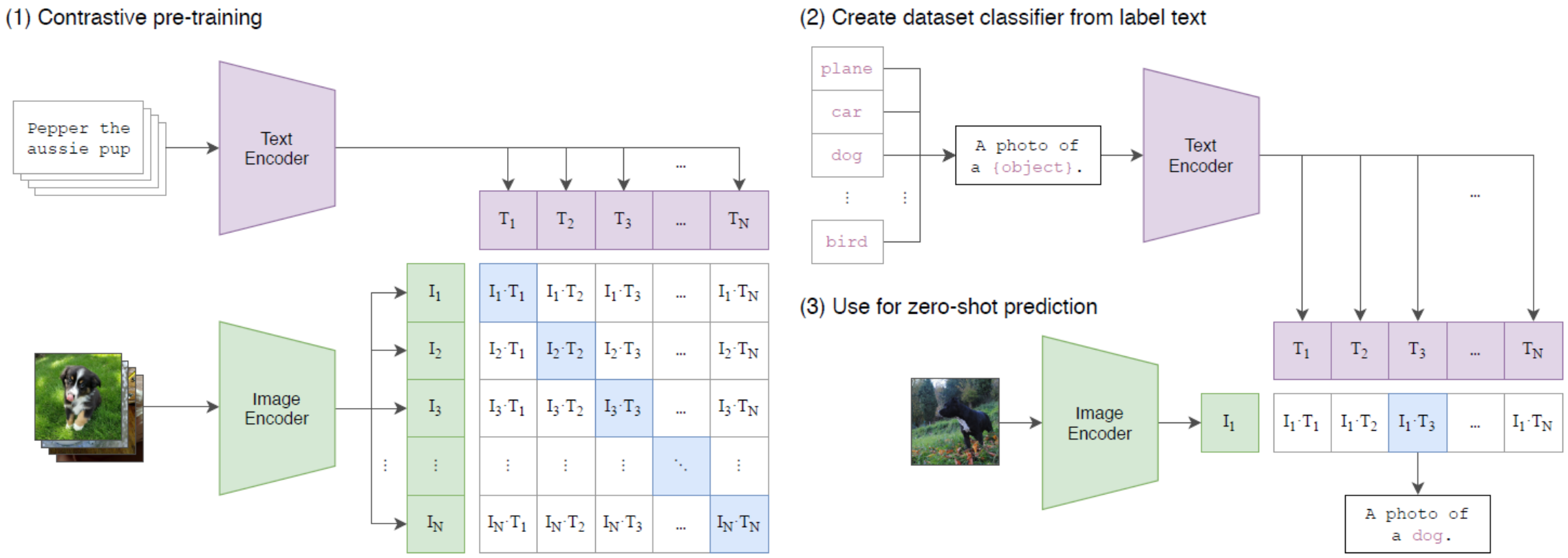
clouds disappeared ☺

High-Resolution Video Synthesis

Outline

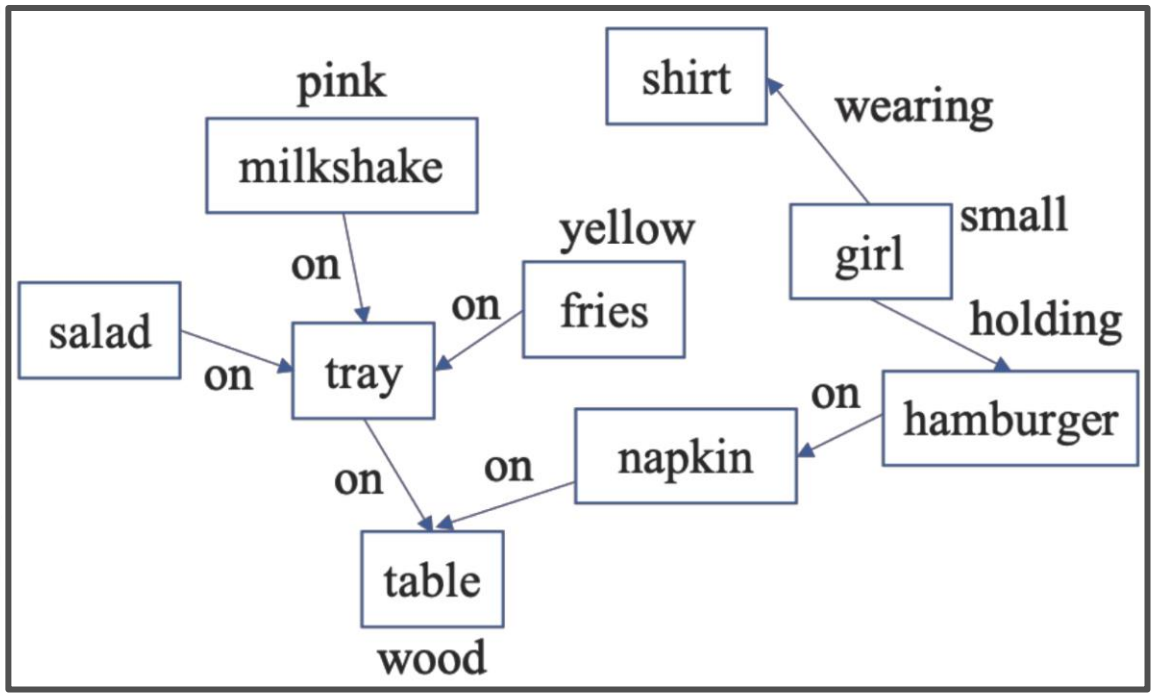
- Language Pre-training
- Vision Pre-training
- Vision-Language Pre-training
- Language-enhanced CV
- Vision-enhanced NLP
- Summarization

Text Supervision-based Enhancement (for visual understanding)



Alec Radford et al. **Learning Transferable Visual Models From Natural Language Supervision**. ICML 2021.

Text Expansion-based Enhancement (for visual reasoning)



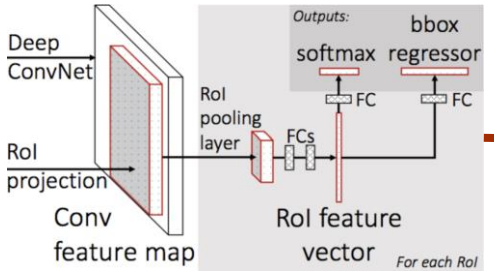
What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl, holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food, on → **Choose** color: yellow | brown

Hudson and Manning. **GQA: a New Dataset for Compositional Question Answering over Real-World Images**. 2019.

Dynamic Context Expansion

1 detected objects



Faster R-CNN

girl 0.91; **table** 0.87;
salad 0.73; **fries** 0.95;
hamburger 0.74;
beverage 0.69; ...

context expansion by common sense

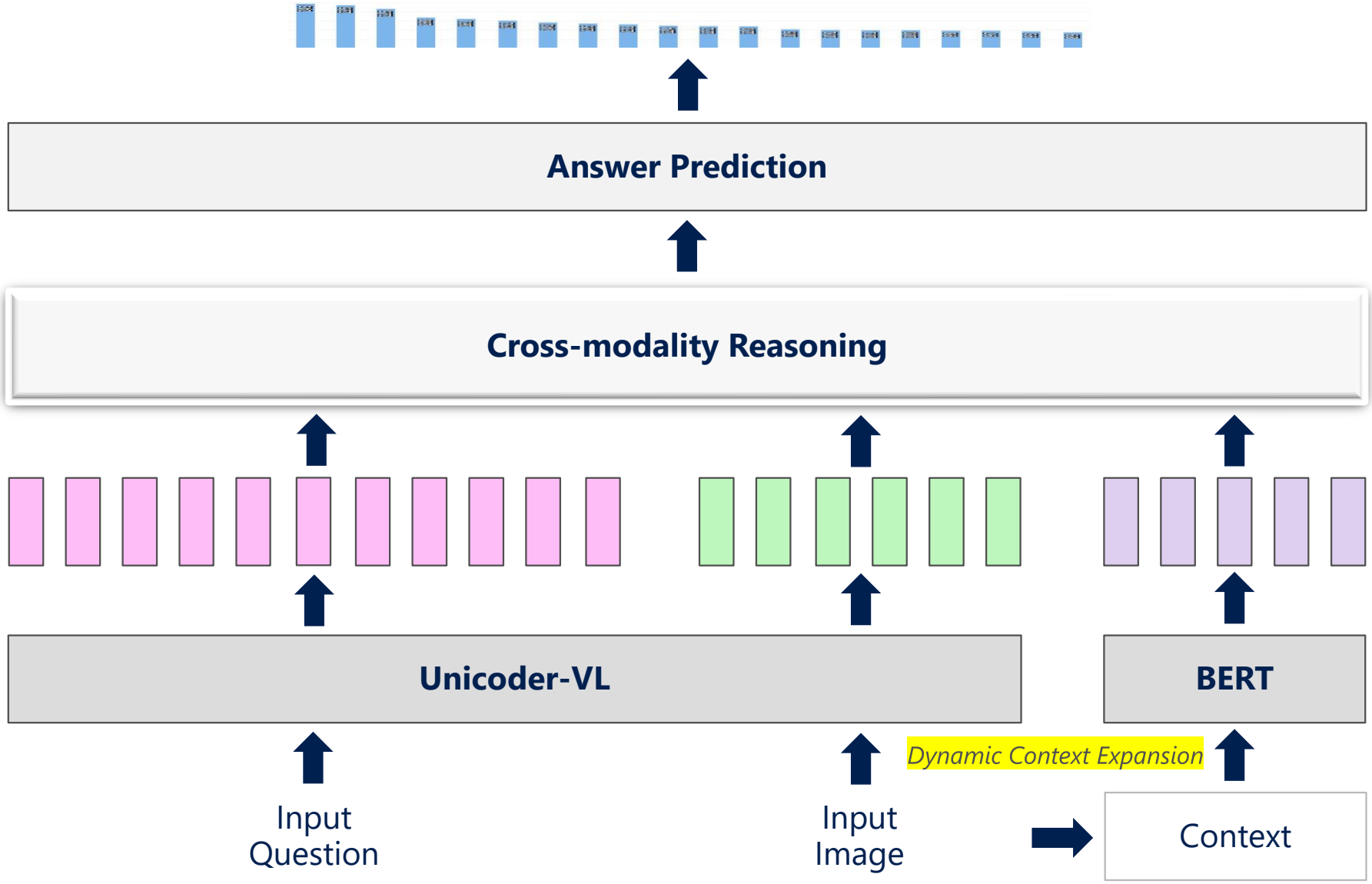
2 expanded triples

table is wood
fries is yellow
hamburger contains vegetables
napkin is white
beverage can drink
...

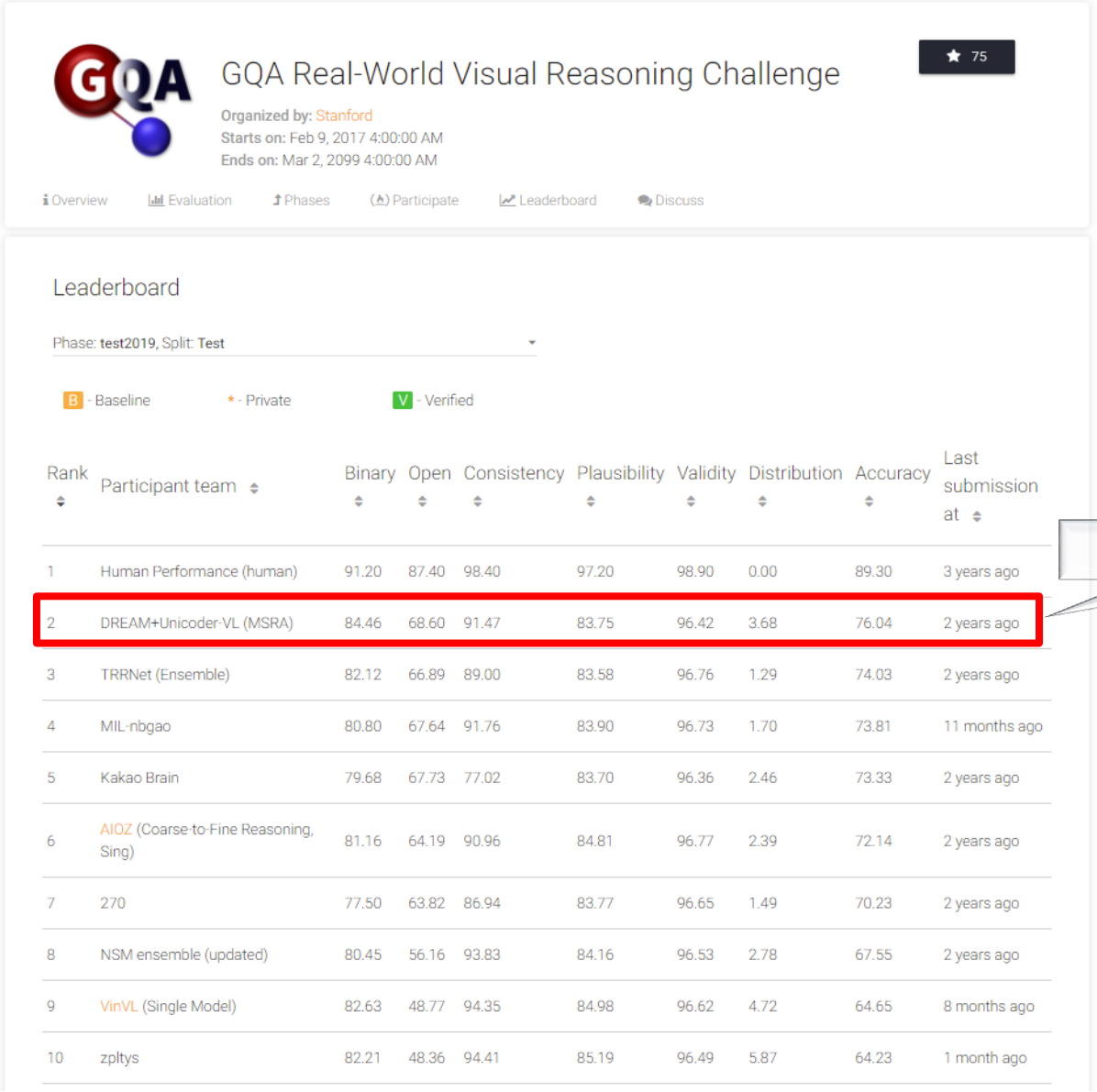
- **what color are the chips?**
- **what is located to the left of the fries?**
- **where is the tray located in?**
- **what color is the thing under the food left of the little girl with the yellow shirt?**
- **...**

3 co-click queries of the same image

Text Expansion-based Visual Reasoning



State-of-the-Art on the GQA Leaderboard



GQA Real-World Visual Reasoning Challenge

★ 75

Organized by: Stanford
Starts on: Feb 9, 2017 4:00:00 AM
Ends on: Mar 2, 2099 4:00:00 AM

Overview | Evaluation | Phases | Participate | Leaderboard | Discuss

Leaderboard

Phase: test2019, Split: Test

B - Baseline * - Private V - Verified

Rank	Participant team	Binary	Open	Consistency	Plausibility	Validity	Distribution	Accuracy	Last submission at
1	Human Performance (human)	91.20	87.40	98.40	97.20	98.90	0.00	89.30	3 years ago
2	DREAM+Unicoder-VL (MSRA)	84.46	68.60	91.47	83.75	96.42	3.68	76.04	2 years ago
3	TRRNet (Ensemble)	82.12	66.89	89.00	83.58	96.76	1.29	74.03	2 years ago
4	MIL-nbgao	80.80	67.64	91.76	83.90	96.73	1.70	73.81	11 months ago
5	Kakao Brain	79.68	67.73	77.02	83.70	96.36	2.46	73.33	2 years ago
6	AIOZ (Coarse-to-Fine Reasoning, Sing)	81.16	64.19	90.96	84.81	96.77	2.39	72.14	2 years ago
7	270	77.50	63.82	86.94	83.77	96.65	1.49	70.23	2 years ago
8	NSM ensemble (updated)	80.45	56.16	93.83	84.16	96.53	2.78	67.55	2 years ago
9	VinVL (Single Model)	82.63	48.77	94.35	84.98	96.62	4.72	64.65	8 months ago
10	zpltys	82.21	48.36	94.41	85.19	96.49	5.87	64.23	1 month ago

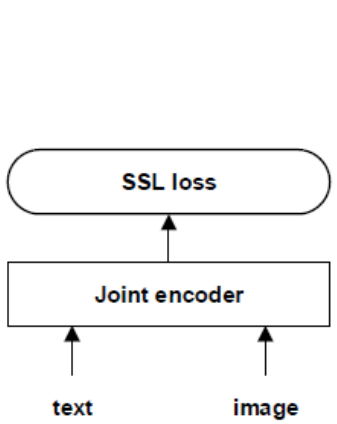
1/115 (as of 2021-08)

https://eval.ai/web/challenges/challenge-page/225/leaderboard/733

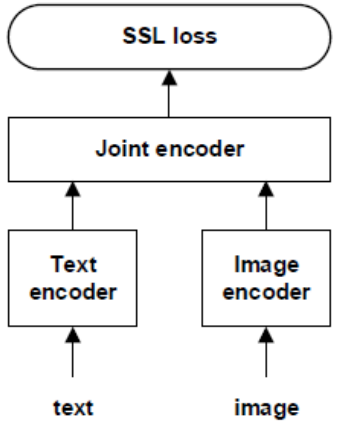
Outline

- Language Pre-training
- Vision Pre-training
- Vision-Language Pre-training
- Language-enhanced CV
- Vision-enhanced NLP
- Summarization

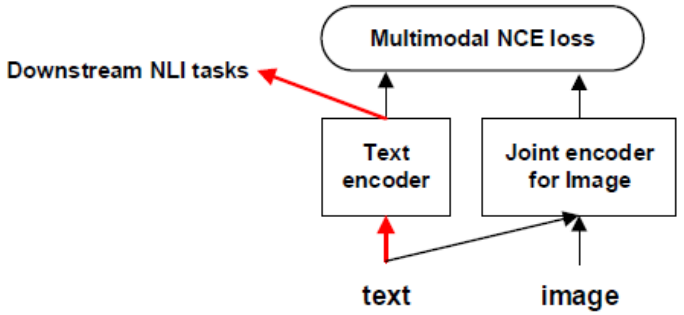
Contrastive Learning-based Enhancement: MACD



(a) Multimodal SSL with one joint encoder.



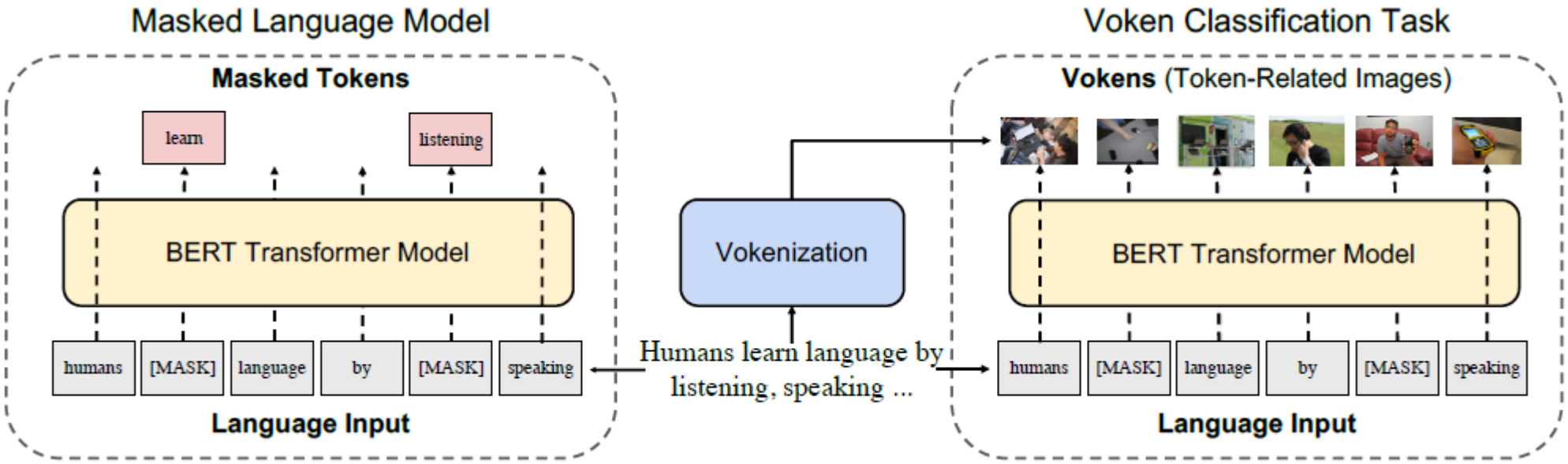
(b) Multimodal SSL with two single-modal encoders and one joint encoder.



(c) Our proposed multimodal aligned contrastive decoupled network. When adapting to downstream NLI tasks, we directly leverage the representation by the text encoder through the red lines, which only requires text as input.

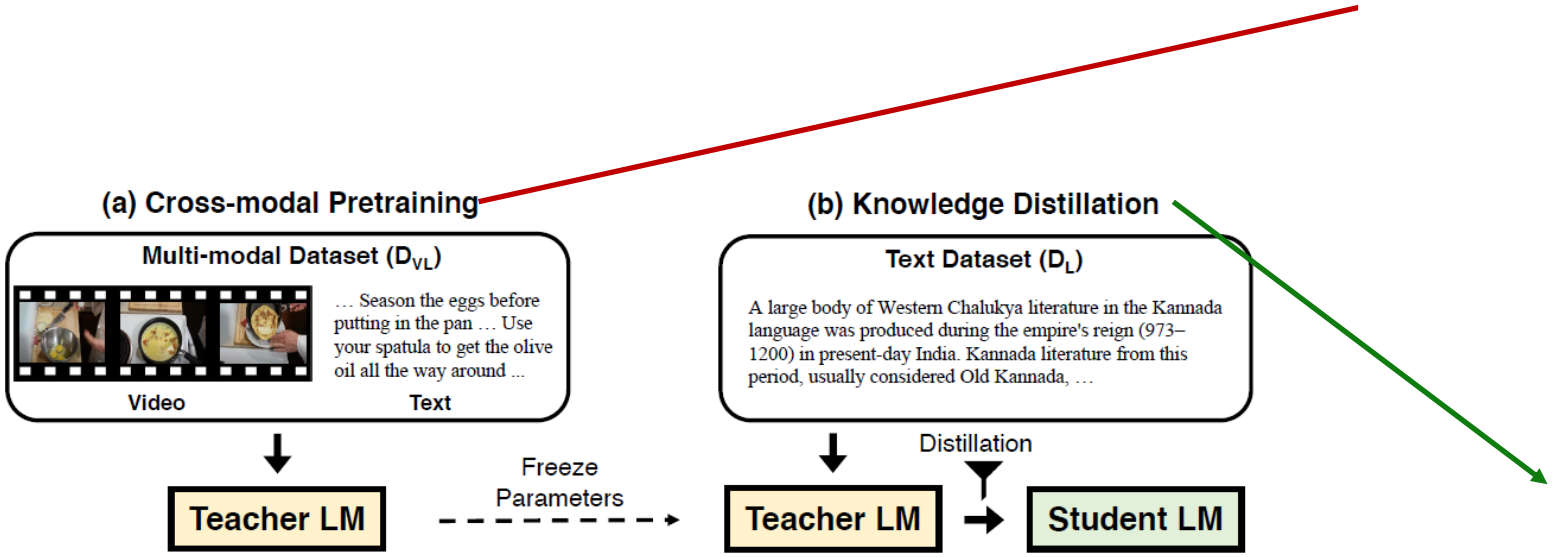
Wanyun Cui, Guangyu Zheng, Wei Wang. **Unsupervised Natural Language Inference via Decoupled Multimodal Contrastive Learning**. EMNLP 2020.

Visual Supervision-based Enhancement: Vokenization



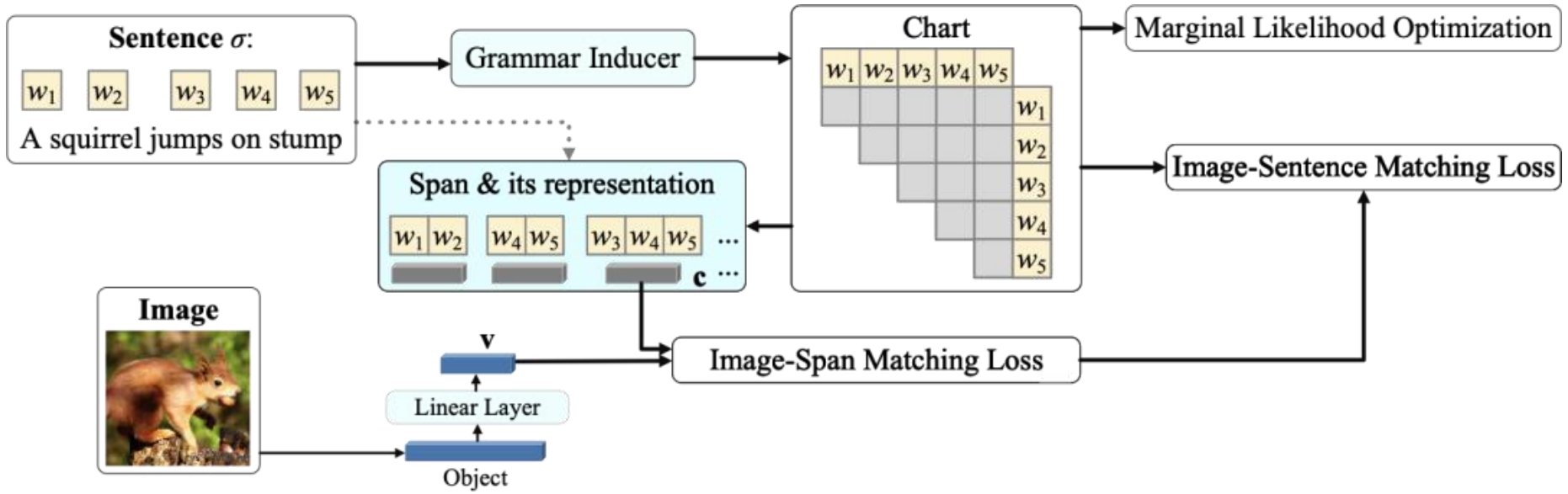
Hao Tan, Mohit Bansal. **Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision**. EMNLP 2020.

Knowledge Distillation-based Enhancement: VIDLANKD



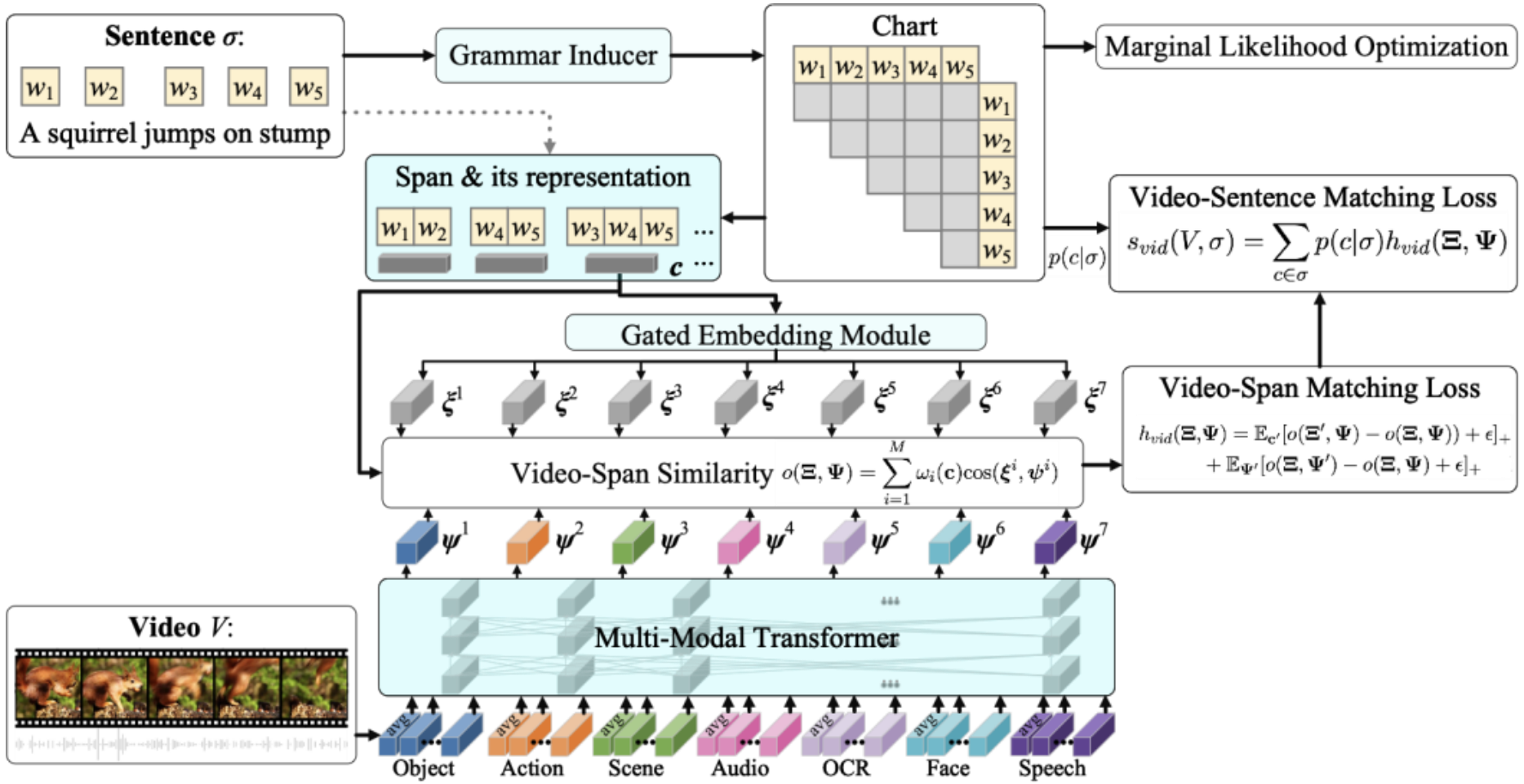
Zineng Tang, Jaemin Cho, Hao Tan, Mohit Bansal. **VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer**. arXiv 2021.

Visual Regularization-based Enhancement: VC-PCFG



Yanpeng Zhao, Ivan Titov. **Visually Grounded Compound PCFGs**. EMNLP 2020.

Visual Regularization-based Enhancement: MMC-PCFG



Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, Jiebo Luo. **Video-aided Unsupervised Grammar Induction**. NNACL 2021.

Outline

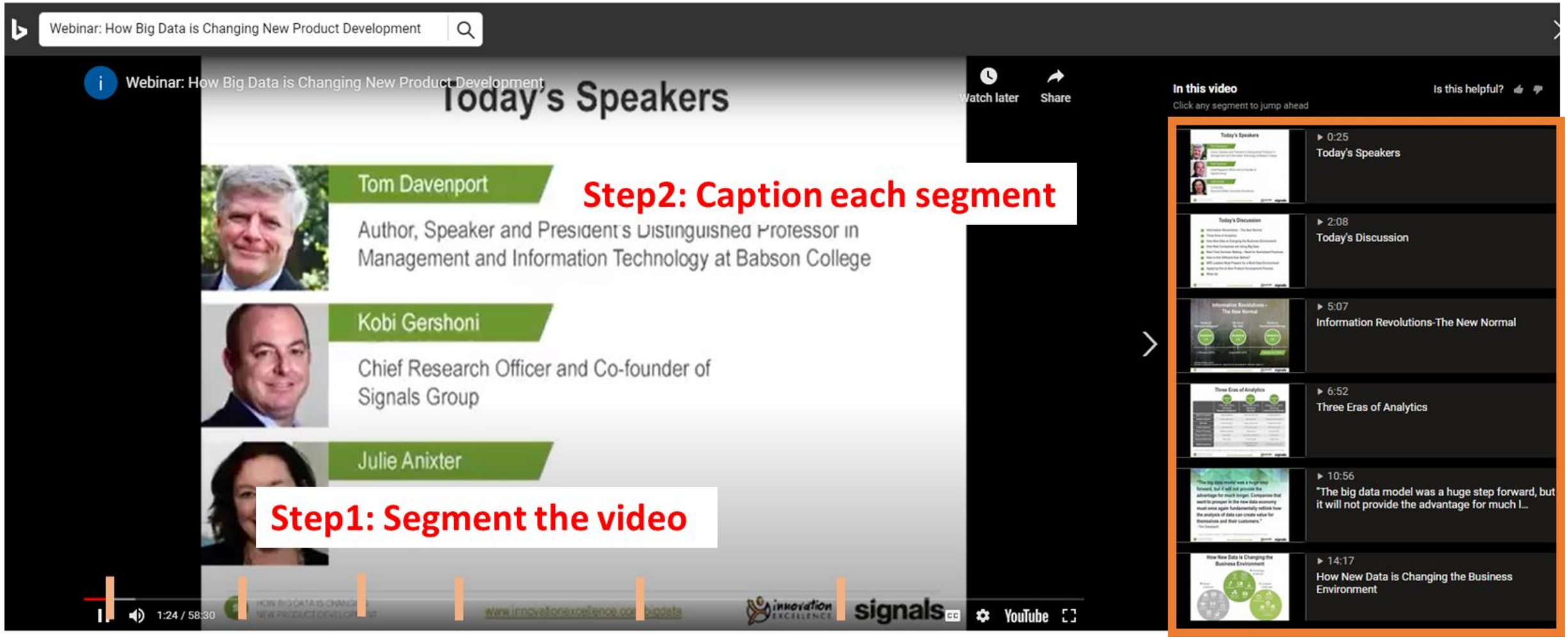
- Language Pre-training
- Vision Pre-training
- Vision-Language Pre-training
- Language-enhanced CV
- Vision-enhanced NLP
- Summarization

Practical Applications

Image Search

Image Captioning

Practical Applications

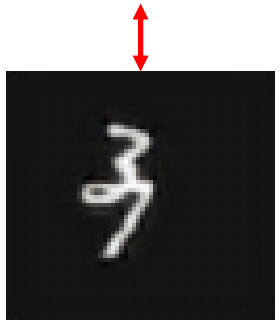


Video Search, Segmentation and Captioning

Big Opportunities Ahead

A baseball game is played.

Digit 7 moves right then left while digit 3 moves down then up.



free transformations between texts, images and videos

[MUM: A new AI milestone for understanding information (blog.google)]

multimodal search, question answering and dialogue

[Multimodal Neurons in Artificial Neural Networks (openai.com)]

model visualization with natural language and vision

Summary

- **NLP and CV trend to be unified.**
  - Similar backbones
  - Similar pre-training tasks
  - Similar representation formats: textual or visual tokens
  - Text as additional signal/supervision for vision understanding (e.g., CLIP)
  - Vision as additional signal/supervision for text understanding (e.g., MMC-PCFG)
- **VL pre-trained models can achieve SOTA results on most existing VL tasks.**
  - Image/Video retrieval
  - Image/Video captioning
  - Image/Video qa & reasoning
- **High-quality VL pairs for VL pre-training are still limited.**
  - Unlimited text, image and video corpus, limited image-text pairs and very limited video-text pairs
- **Visual content generation is hard to evaluation.**
  - No metric or benchmark dataset available for image/video generation.

Challenge

- How to model the spatial-temporal information in image and video?
- How to define visual concepts for vision-language pre-trained models?
- How to leverage unlimited language and vision corpus to enhance vision-language pre-training?
- How to build reasonable and comprehensive vision-language benchmarks?
- How to design evaluation metrics for image/video generation tasks?
- How CV can help NLP better, visual commonsense representation learning?
- How can vision-language pre-trained models help neural model visualization and interpretation?
- What are the next killer apps with vision-language pre-trained models?

To enable NLP systems to look, listen, comprehend and reply!

Thank you!