COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

NUS National University of Singapore

WED-PM-238

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Towards Fast Adaptation of Pretrained Contrastive Models for Multi-Channel Video-Language Retrieval

Xudong Lin[1], Simran Tiwari[1], Shiyuan Huang[1], Manling Li[2],

Mike Zheng Shou[3], Heng Ji[2], Shih-Fu Chang[1]

[1]Columbia University [2]UIUC [3]NUS

Tired of costly multimodal pretext training?

Modality-specific Pretraining

ResNet

...

ViT

Cat

"The cat is a domestic species"

BERT

...

GPT

"... of small carnivorous mammal."

Multimodal Pretext Training

Pretext Tasks:
Vision/Text/Audio Matching
Masked Modeling
...

Text Component

Projection Layers

Vision Component

Large-scale Multimodal Data

Finetuning on Downstream Tasks

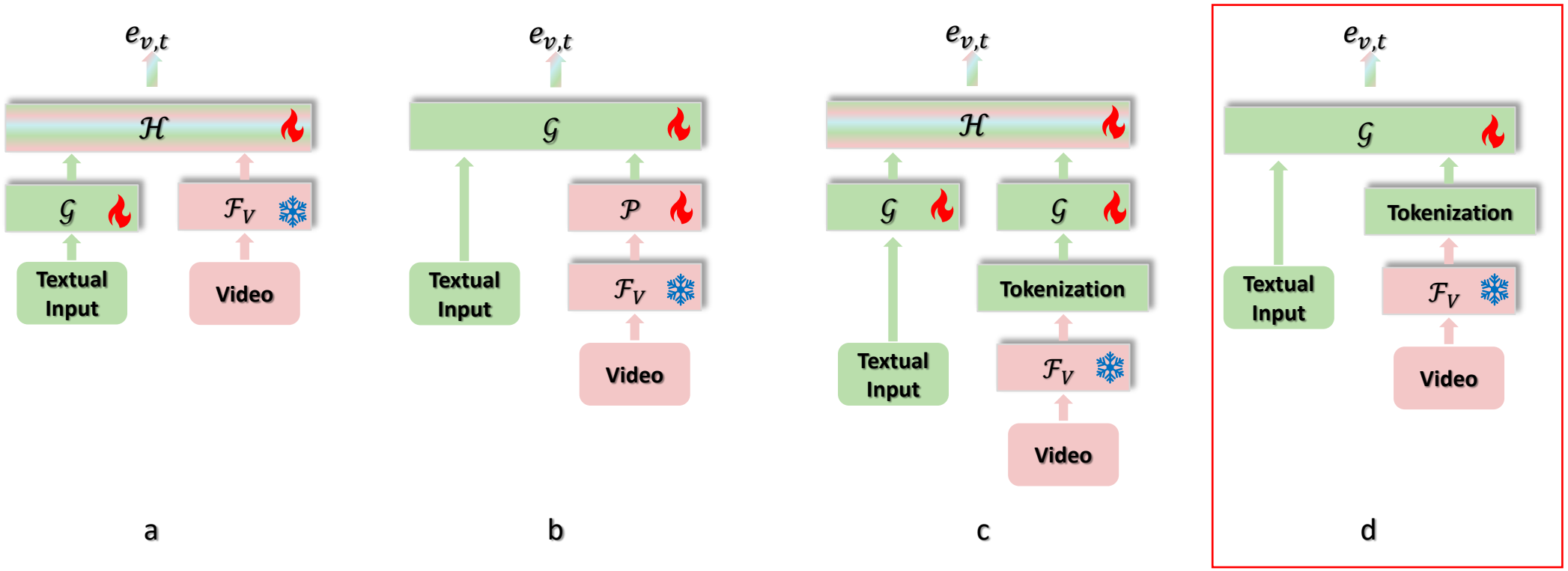
Finetuning

Multimodal Model

Small-scale Data with Annotation

Towards Fast Adaptation of Pretrained Contrastive Models for Multi-Channel Video-Language Retrieval

Video-language learning without costly large-scale pretext training!



$\mathcal{F}_V$ Pretrained Video Encoder

$\mathcal{G}$ Pretrained Text Model

$\mathcal{H}$ Multimodal Transformer

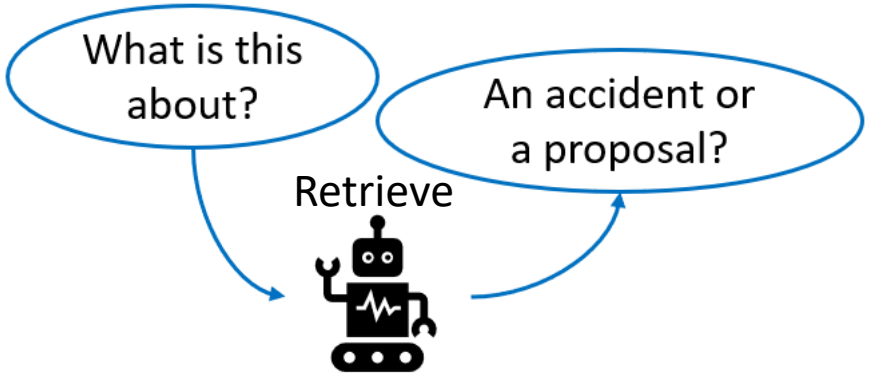
$\mathcal{P}$ Projector

Tokenization

$\mathcal{F}_V(v)$ — Retrieve — $\{\mathcal{F}_T(w_1), \mathcal{F}_T(w_2), \ldots, \mathcal{F}_T(w_{|\mathbb{W}|})\}$

Multi-Channel Video-Language Retrieval

- What is Multi-Channel Video-Language Retrieval?
  - Given a multi-channel sample (e.g., video + text), retrieve a text response, or vice versa.
  - For example:
    - Open-ended video question answering
    - Text-video retrieval
- Key requirements:
  - Fusing multi-channel information
  - Discriminative representations



https://www.youtube.com/watch?v=hya9xxn7CA0

Existing Paradigm

Modality-specific Pretraining

ResNet

...

ViT

Cat

"The cat is a domestic species"

BERT

...

GPT

"... of small carnivorous mammal."

Multimodal Pretext Training

Pretext Tasks:
Vision/Text/Audio Matching
Masked Modeling
...

Text Component

Projection Layers

Vision Component

Large-scale Multimodal Data

Finetuning on Downstream Tasks

Finetuning

Multimodal Model

Small-scale Data with Annotation

Research Question: Can we remove the costly multimodal pretext training stage?

Modality-specific Pretraining

ResNet

...

ViT

Cat

"The cat is a domestic species"

BERT

...

GPT

"... of small carnivorous mammal."

~~Multimodal Pretext Training~~

Pretext Tasks:
Vision/Text/Audio Matching
Masked Modeling
...

Text Component

Projection Layers

Vision Component

Large-scale Multimodal Data

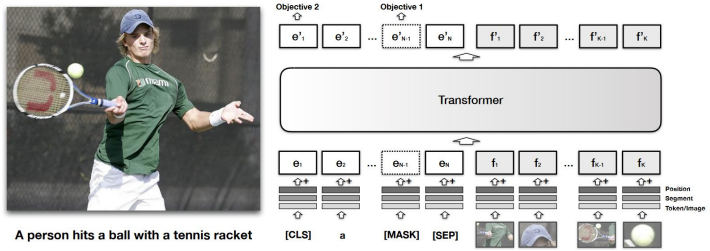
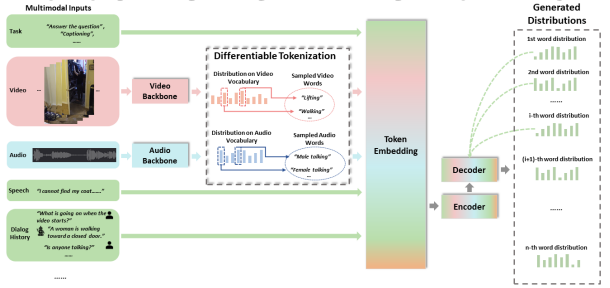
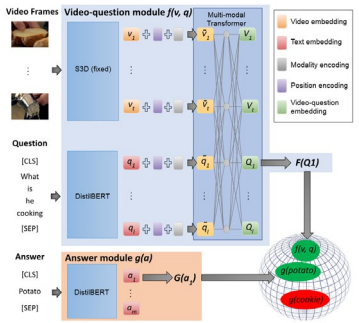
Finetuning on Downstream Tasks

Finetuning

Multimodal Model

Small-scale Data with Annotation

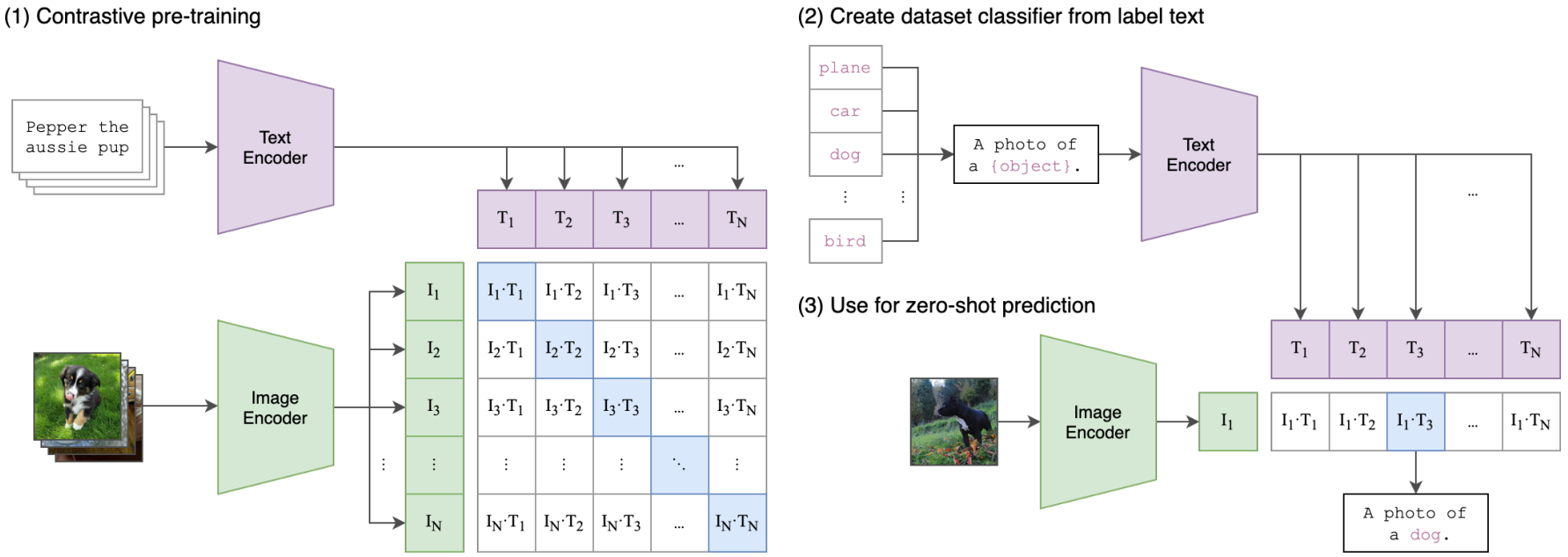
Design Space of Multimodal Models

Model Design Space	What is the representation of non-textual modalities, Continuous features or text tokens?	
What is the fusion strategy, Text Transformer Or Multimodal Transformer?	Continuous + Text Trans.	Text tokens + Text Trans.
	 (Li, Liunian Harold, et al., 2019)	 (Lin, Xudong, et al., 2021)
	Continuous + MM Trans.	Text tokens + MM Trans.
	 (Yang, Antoine, et al., 2021)	

Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." *arXiv preprint arXiv:1908.03557* (2019).
Lin, Xudong, et al. "Vx2Text: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
Yang, Antoine, et al. "Just ask: Learning to answer questions from millions of narrated videos." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

Background: Multimodal Contrastive Models

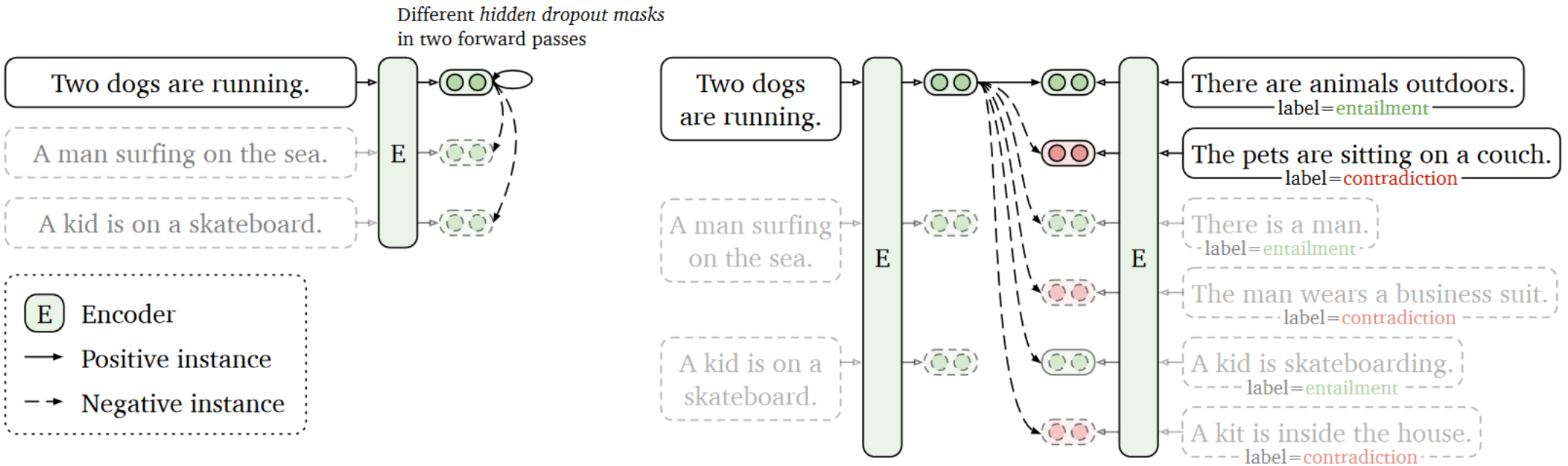
- Assumption: we have access to multimodal contrastive models like CLIP (Radford, Alec, et al., 2021) to perform zero-shot visual concept extraction.



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

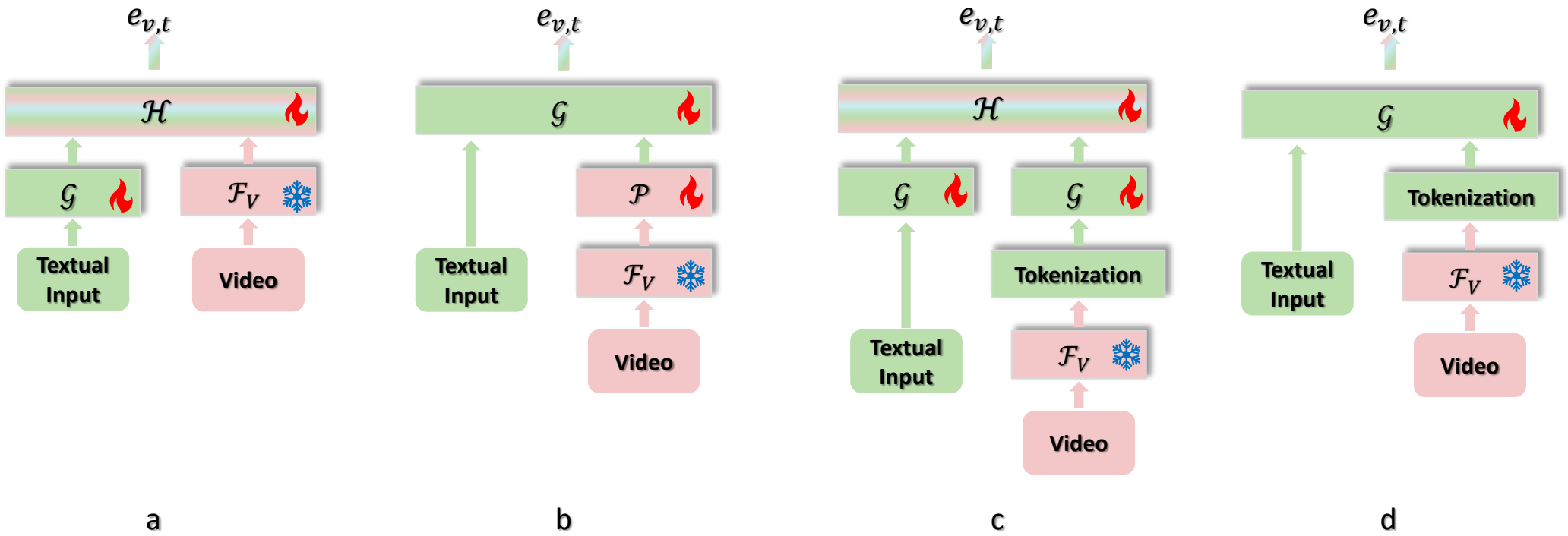
Background: Contrastive Text Model

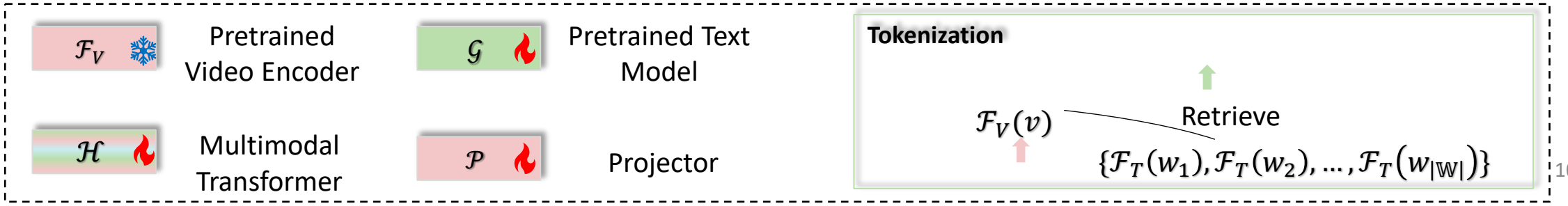
- Assumption: we have access to pretrained contrastive text models like SBERT/SimCSE to produce discriminative representations.



Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).
Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." *arXiv preprint arXiv:2104.08821* (2021).

Towards Fast Adaptation of Pretrained Contrastive Models





Experimental Results

- Setting
  - Task type:
    - Open-ended, given question and video, retrieve the most relevant answer from the collection of all answers in the dataset.
    - Multiple choice
    - Text-video retrieval

Model	Open-ended Acc(%)		Multi-Choice Acc(%)	Retrieval AveR (%)	
	iVQA	ActivityNet-QA	How2QA	YouCook II	VATEX
Conti. + Multi.	22.4	36.9	79.2	41.9	69.4
Conti. + Text	23.2	37.3	80.4	46.2	72.7
Text + Multi.	23.4	37.1	79.4	40.4	67.5
Text + Text	**31.6**	**38.7**	**82.9**	**49.7**	**74.8**

Experimental Results

- Setting
  - Task type:
    - Open-ended, given question and video, retrieve the most relevant answer from the collection of all answers in the dataset.
    - Multiple choice
    - Text-video retrieval

Model	Open-ended Acc(%)		Multi-Choice Acc(%)	Retrieval AveR (%)	
	iVQA	ActivityNet-QA	How2QA	YouCook II	VATEX
Conti. + Multi.	22.4	36.9	79.2	41.9	69.4
Conti. + Text	23.2	37.3	80.4	46.2	72.7
Text + Multi.	23.4	37.1	79.4	40.4	67.5
Text + Text	**31.6**	**38.7**	**82.9**	**49.7**	**74.8**

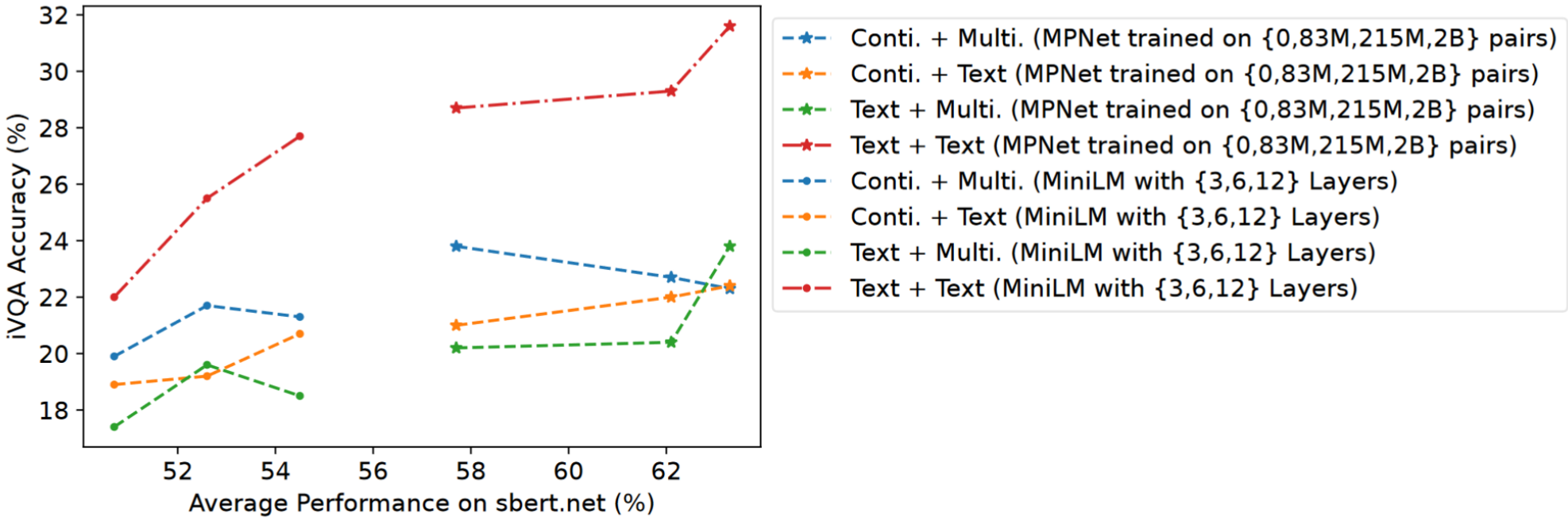
The Text + Text variant consistently achieves superior performance!

Comparison to SOTA

- Comparable or even better performance without extra multimodal pretext samples for training!
- Easily upgraded w.r.t. multimodal contrastive models!

Model	$\mathcal{F}_V, \mathcal{F}_T$	Extra MM Samples	Δ GPU hours	iVQA	ActivityNet	How2QA
MERLOT [38]	-	180M	-	-	41.4	-
SiaSamRea [36]	-	5.6M + 80K	-	-	39.8	84.1
VQA-T [33]	S3D [15]	69M + 3M	350 + 30	35.2	**39.0**	**85.3**
Conti. + Multi.	S3D [15]	69M	400	35.4	38.9	84.4
Conti. + Multi. (+ *ASR*)	S3D [15]	69M	400	36.0	38.9	84.8
Text + Text (Ours)	S3D [15]	0	0	31.6	38.7	82.9
Text + Text (+ *ASR*, Ours)	S3D [15]	0	0	**36.8**	38.8	84.6
FrozenBiLM [34]	CLIP [19]	10M	160	39.7	43.2	81.5
FrozenBiLM [34] (+ *ASR*)	CLIP [19]	10M	160	39.6	**43.2**	86.7
Text + Text (Ours)	CLIP [19]	0	0	36.9	41.4	92.4
Text + Text (+ *ASR*, Ours)	CLIP [19]	0	0	**40.2**	41.4	**93.2**

Ablation: how much do we gain by using better pretrained textual transformers?



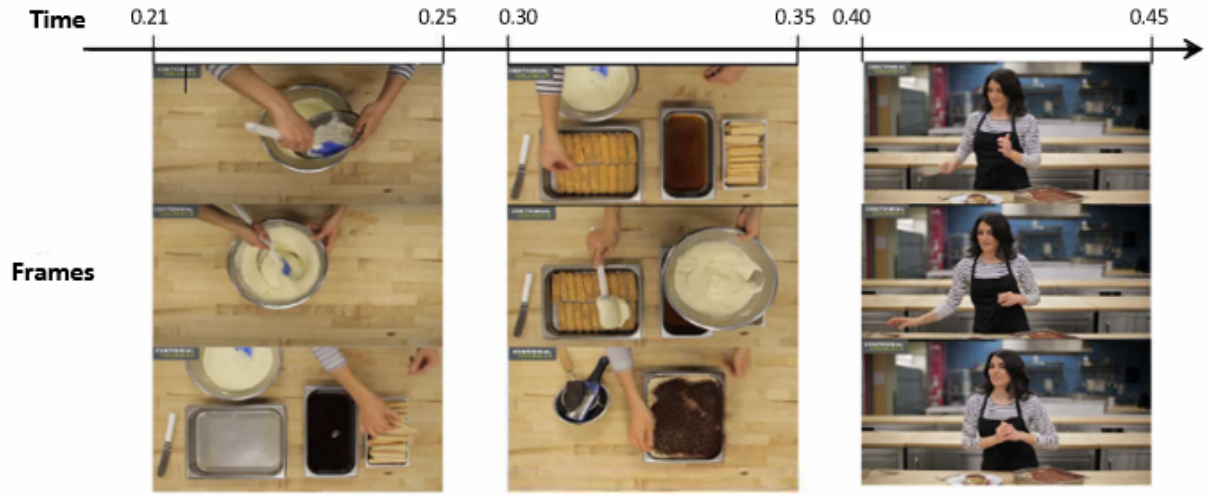
Ablation: what is the best vocabulary?

- Answer words produce the best results.

Vocabulary Source	iVQA Accuracy (%)
60K frequent English words	27.3
Visual Genome Objects and Attributes (VG-OA)	29.3
Words from all answers in iVQA training set (AW)	31.6
AW & VG-OA	29.5

Qualitative Results

- For more than 64% of iVQA samples, retrieved words include the ground-truth answer;
- When not directly retrieving the answer word, the model still learns to reason the correct answer.



Question	What is the woman wearing?
Text Tokens	First sieve. batter. rubber spatula. blueberries. pie tin. sifter. cake pan. cake mix. meringue. ladle. spatula. pie crust. pans. crust. ricotta. Then chocolate syrup. danish oil. chocolate sauce. waffles. custard. chocolate bars. wooden spatula. baker's chocolate. chocolate. dark chocolate. marinade. chocolates. biscuits. rubber spatula. cooking spray. Then waffle iron. slicer. potato pancakes. knife rack. baking pan. brownies. waffles. grater. oven mitts. cake pan. torte. fudge. brownie. chocolate bars. hot fudge.
Predicted Answer	apron
Ground Truth Answers	apron, apron, striped shirt, apron, apron

Take-home messages

- Using text tokens of visual concepts to represent the visual information and using the pretrained text model as the modality fuser, a comparable or even better multimodal model can be obtained without costly multimodal pretraining!
- We hope this sheds light to upgradable multimodal intelligence.

- *We release our code in the repo.* 
- *Welcome to our poster session on Wednesday afternoon.*
- *Thanks!*