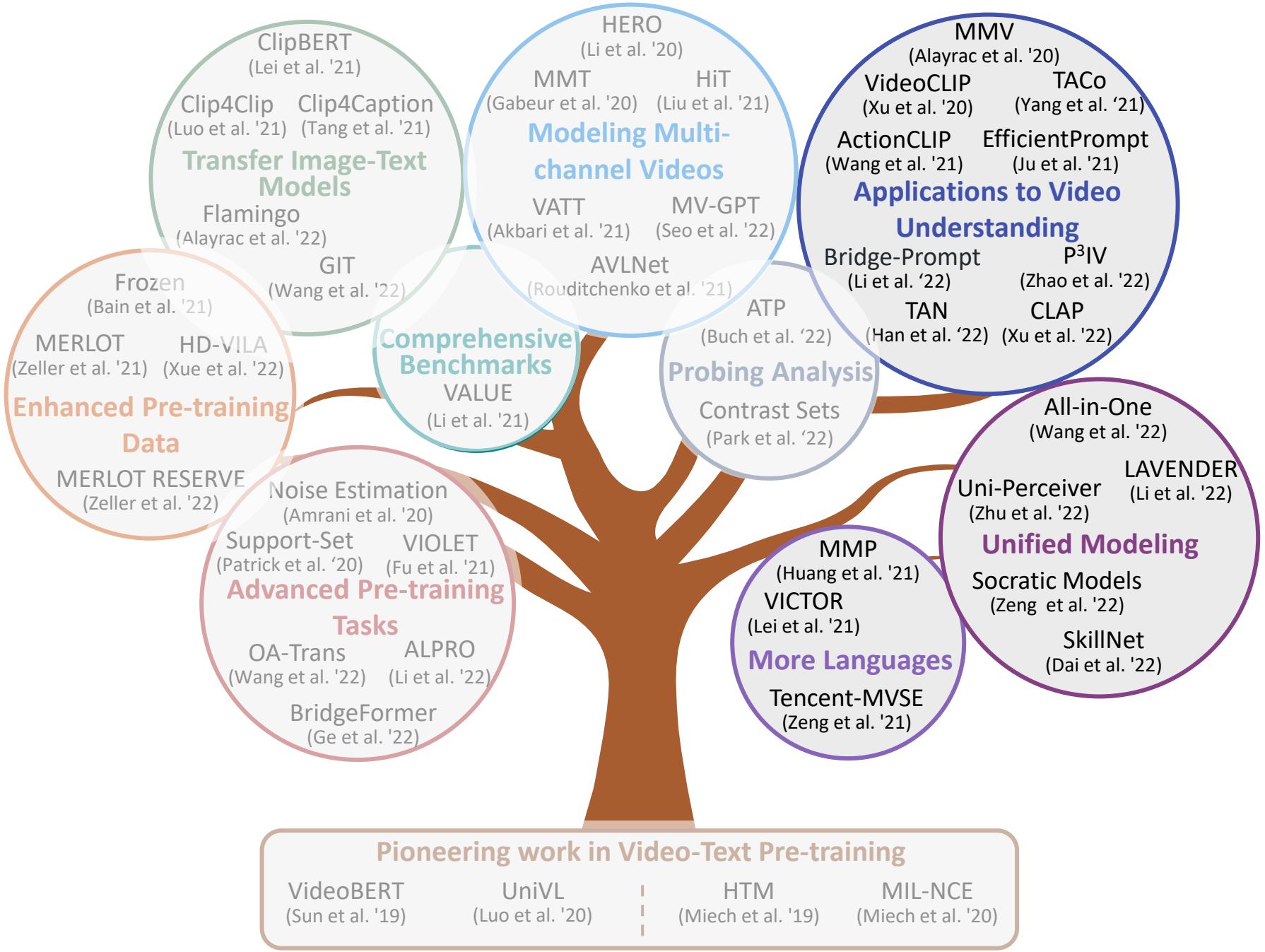
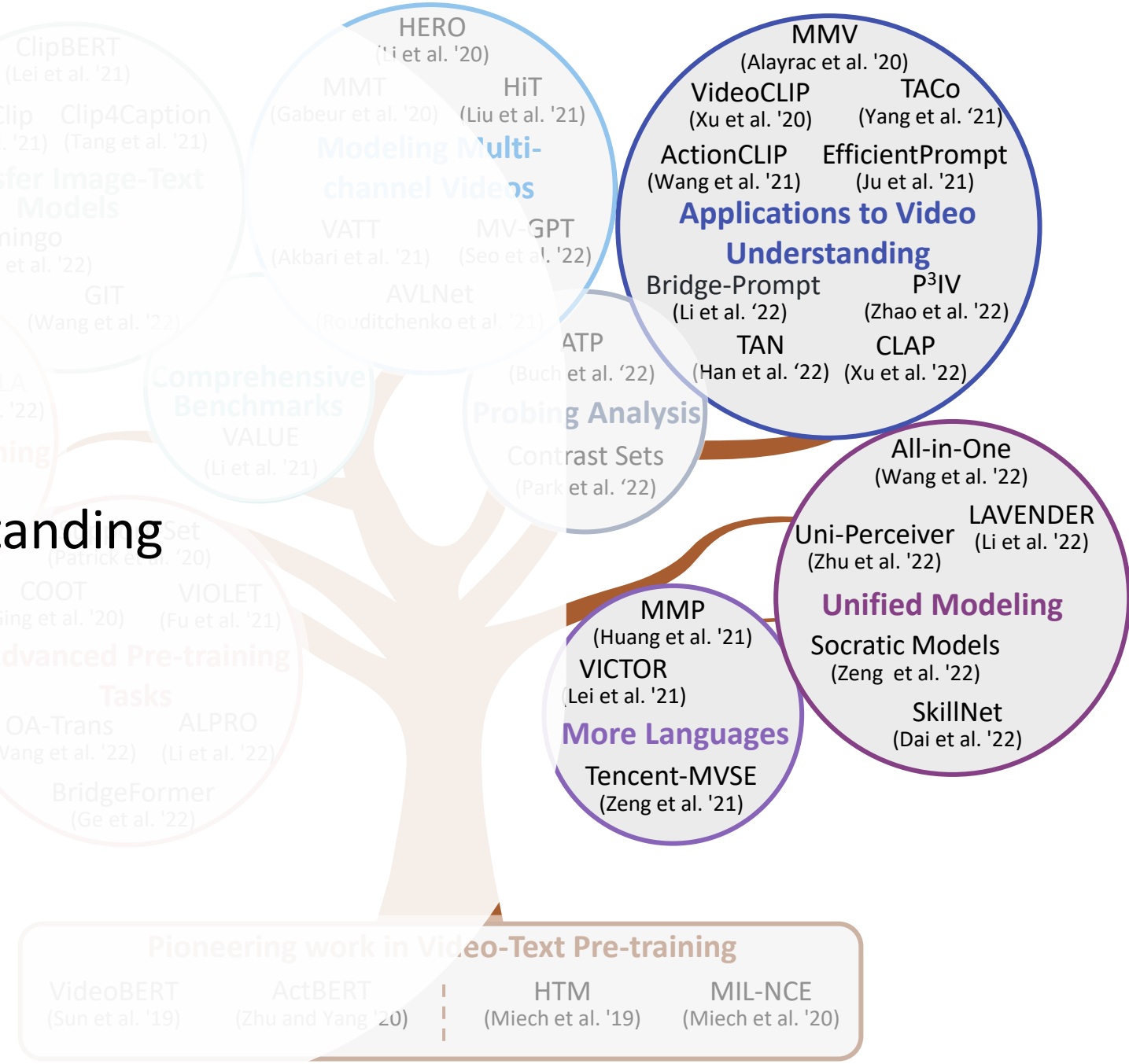
Advanced Topics in Video-Text Pre-training

Chung-Ching Lin


ClipBERT
(Lei et al. '21)
Clip4Clip
(Luo et al. '21)
Clip4Caption
(Tang et al. '21)
Transfer Image-Text Models
Flamingo
(Alayrac et al. '22)
GIT
(Wang et al. '22)
HERO
(Li et al. '20)
MMT
(Gabeur et al. '20)
HiT
(Liu et al. '21)
Modeling Multi-channel Videos
VATT
(Akbari et al. '21)
MV-GPT
(Seo et al. '22)
AVLNet
(Rouditchenko et al. '21)
MMV
(Alayrac et al. '20)
VideoCLIP
(Xu et al. '20)
TACo
(Yang et al. '21)
ActionCLIP
(Wang et al. '21)
EfficientPrompt
(Ju et al. '21)
Applications to Video Understanding
Bridge-Prompt
(Li et al. '22)
P3IV
(Zhao et al. '22)
TAN
(Han et al. '22)
CLAP
(Xu et al. '22)
Frozen
(Bain et al. '21)
MERLOT
(Zeller et al. '21)
HD-VILA
(Xue et al. '22)
Enhanced Pre-training Data
MERLOT RESERVE
(Zeller et al. '22)
Comprehensive Benchmarks
VALUE
(Li et al. '21)
ATP
(Buch et al. '22)
Probing Analysis
Contrast Sets
(Park et al. '22)
All-in-One
(Wang et al. '22)
Uni-Perceiver
(Zhu et al. '22)
LAVENDER
(Li et al. '22)
Unified Modeling
Socratic Models
(Zeng et al. '22)
SkillNet
(Dai et al. '22)
Noise Estimation
(Amrani et al. '20)
Support-Set
(Patrick et al. '20)
VIOLET
(Fu et al. '21)
Advanced Pre-training Tasks
OA-Trans
(Wang et al. '22)
ALPRO
(Li et al. '22)
BridgeFormer
(Ge et al. '22)
MMP
(Huang et al. '21)
VICTOR
(Lei et al. '21)
More Languages
Tencent-MVSE
(Zeng et al. '21)
Pioneering work in Video-Text Pre-training
VideoBERT
(Sun et al. '19)
UniVL
(Luo et al. '20)
HTM
(Miech et al. '19)
MIL-NCE
(Miech et al. '20)

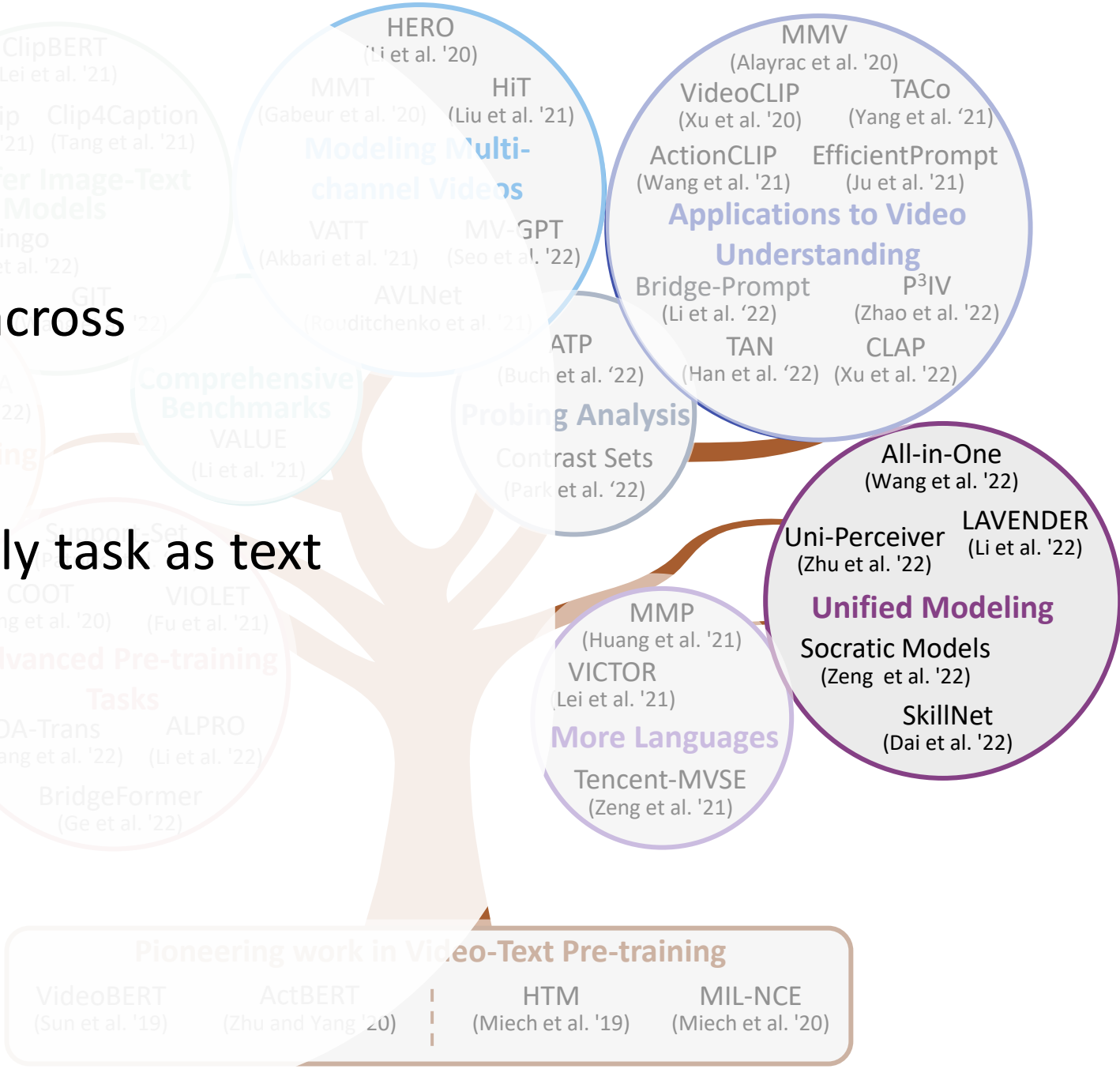
Agenda

- Unified modeling
  - Modality-agnostic
  - Task-agnostic
- Applications to video understanding
  - Video-text contrastive models
  - Transfer image-text models
  - Prompt tuning
- More languages



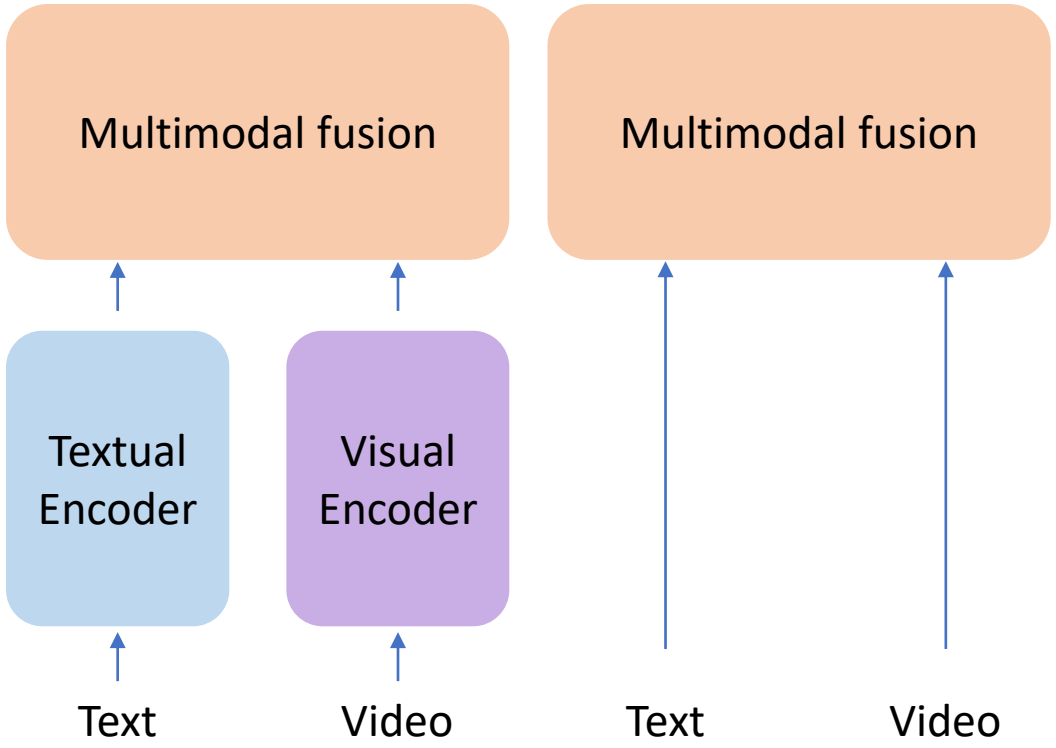
Unification

- Modalities: unify backbones across modalities
- Tasks: unify VL tasks and V-only task as text generation



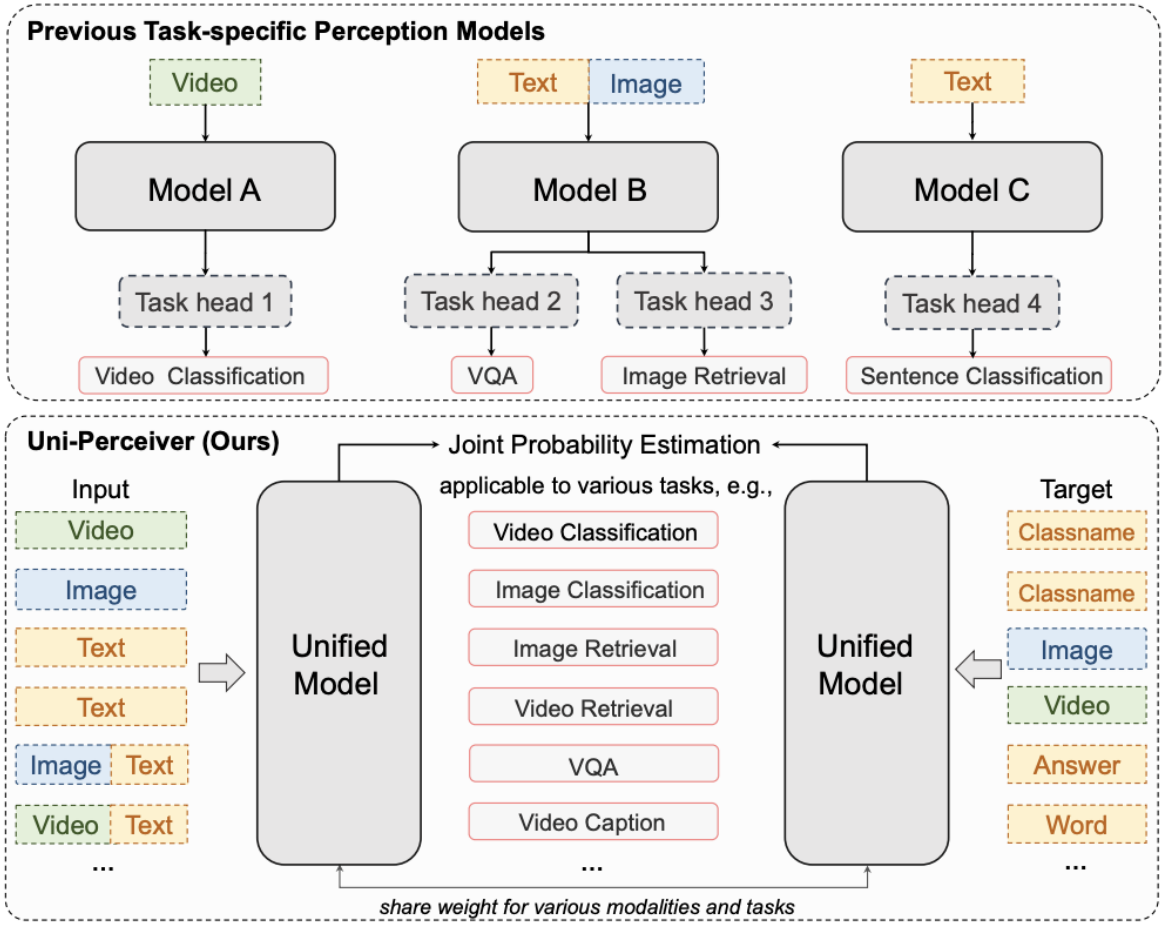
Unification

- Modalities: unify backbones (text, image, video, audio)
  - Uni-Perceiver, All in One, SkillNet, VATT
- Tasks: unify VL tasks V-only task as text generation
  - LAVENDER, Socratic Models, GIT, Flamingo



Modality-specific to modality-agnostic

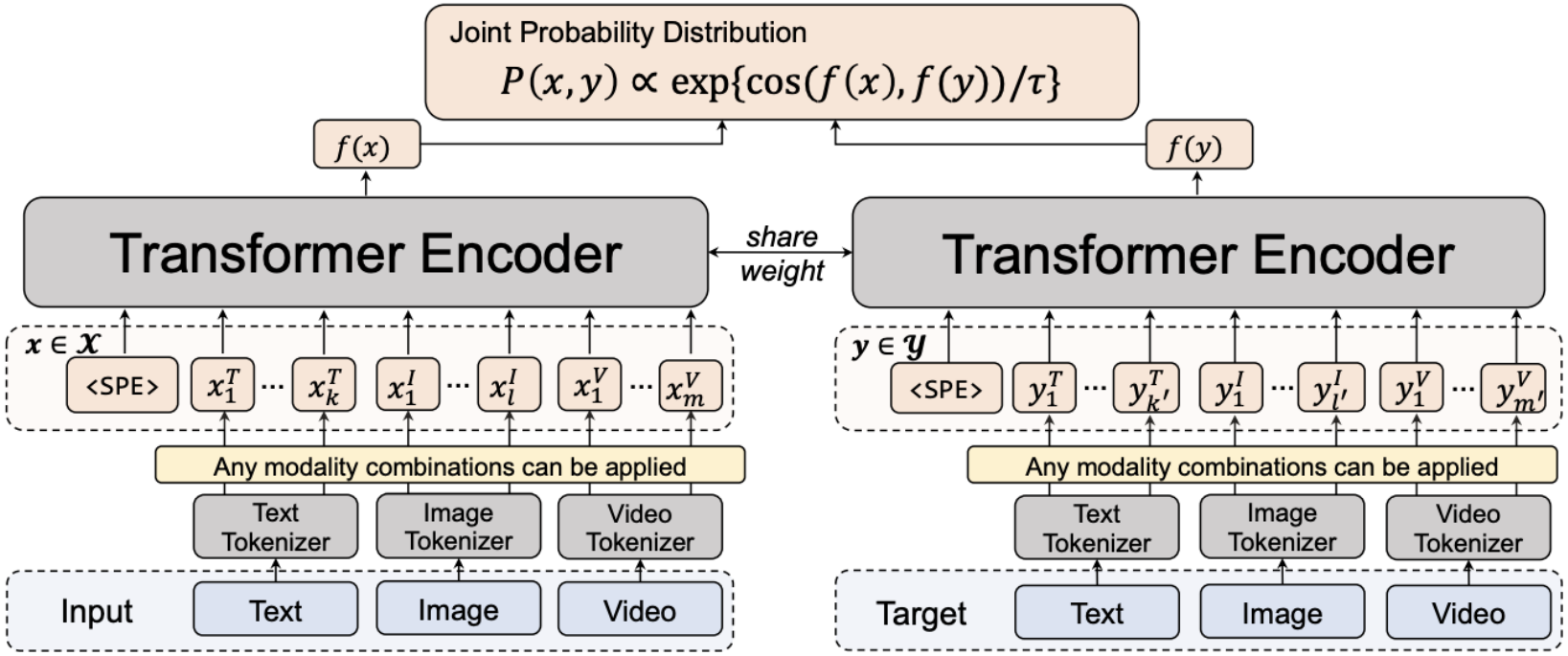
- A modality-agnostic Transformer encoder and lightweight modality-specific tokenizer
- Encodes different task inputs and targets from arbitrary modalities into a unified representation

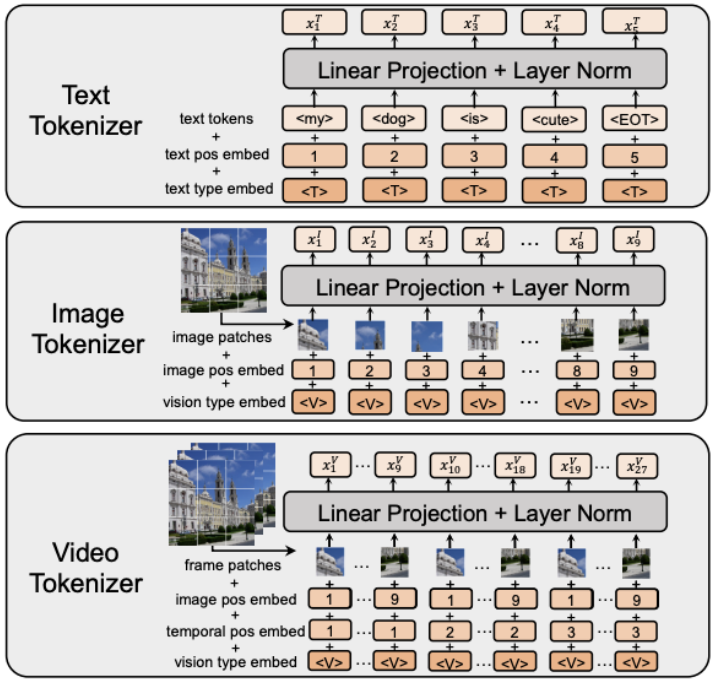


Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks, CVPR 2022

Modality-agnostic model

- Encodes different task inputs and targets from arbitrary modalities into a unified representation



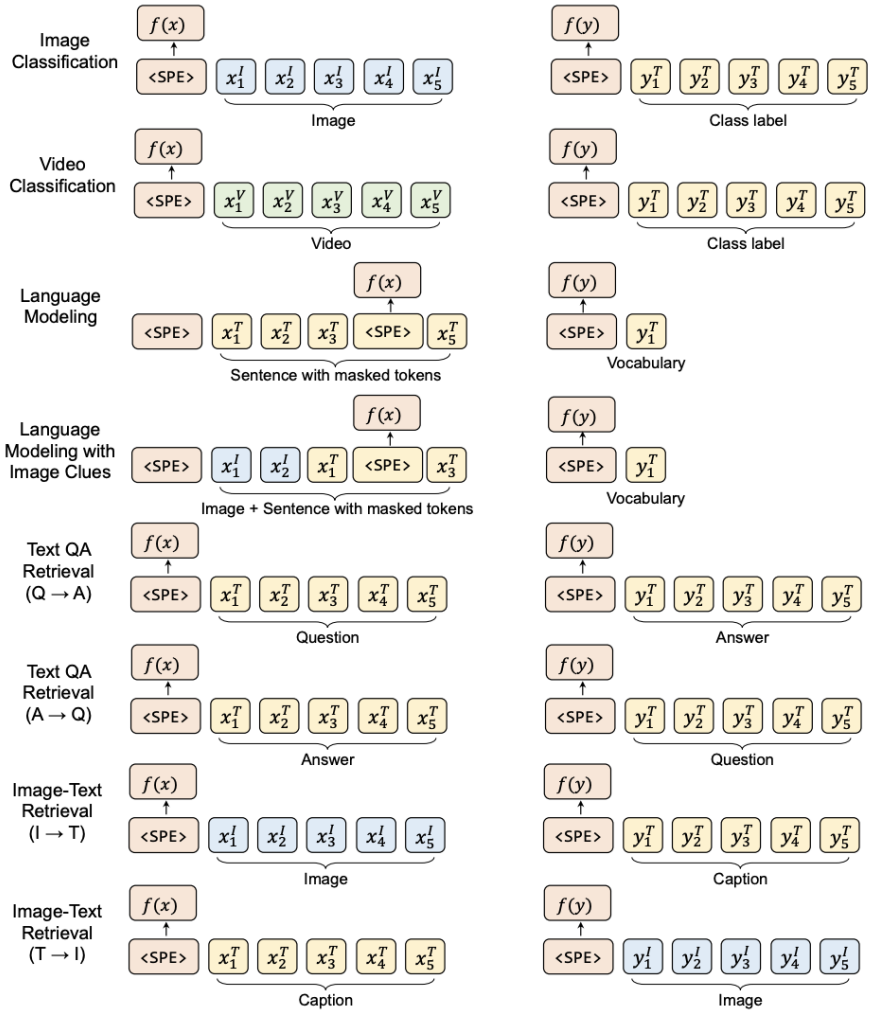


Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks, CVPR 2022

Pre-training on Multi-Modal Tasks

Pre-training dataset

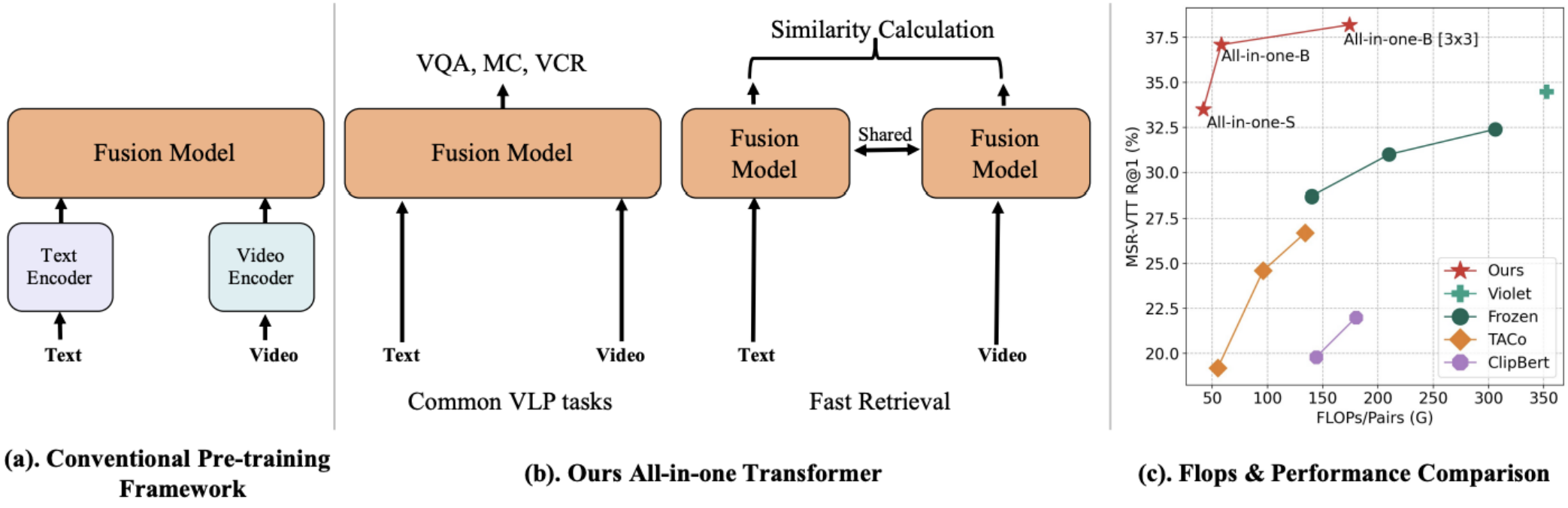
Dataset	#Images	#Videos	#Text
ImageNet-21k [17]	14.2M	0	21K
Kinetics-700 [32]	0	542K	700
Moments in Time [57]	0	792K	339
Books&Wiki [93]	0	0	101M
PAQ [41]	0	0	65M
CC3M [68]	3.0M	0	3.0M
CC12M [9]	11.1M	0	11.1M
COCO Caption [12]	113K	0	567K
Visual Genome [36]	108K	0	5.41M
SBU [58]	830K	0	830K
YFCC [31]	14.8M	0	14.8M



Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks, CVPR 2022

Modality-Agnostic Video-Language Pre-training models

- An unified backbone architecture



(a). Conventional Pre-training Framework

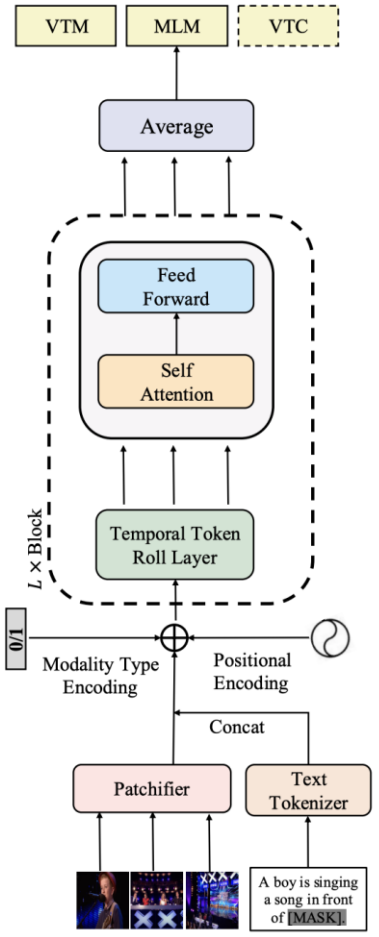
(b). Ours All-in-one Transformer

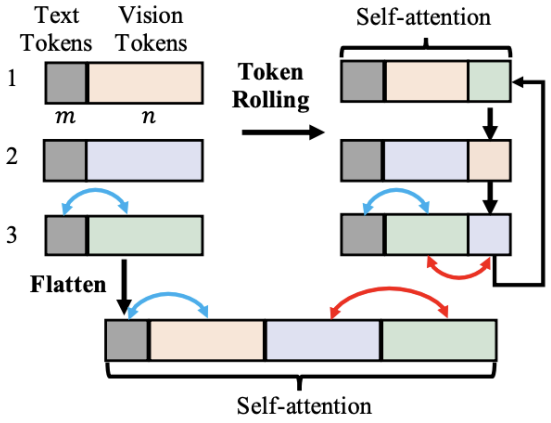
(c). Flops & Performance Comparison

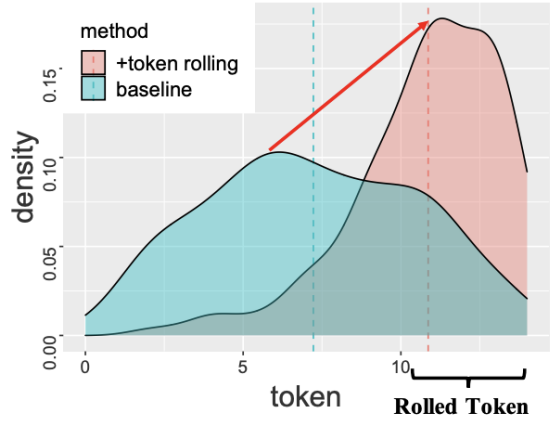
All in One: Exploring Unified Video-Language Pre-training, arXiv 2022

Modality-Agnostic Video-Language Pre-training models

- Token rolling operation to encode temporal representations from video clips in a non-parametric manner



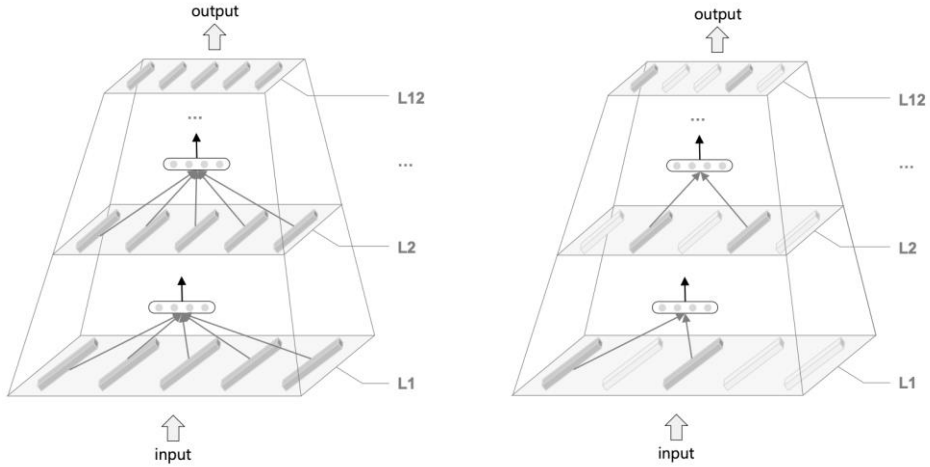




All in One: Exploring Unified Video-Language Pre-training, arXiv 2022

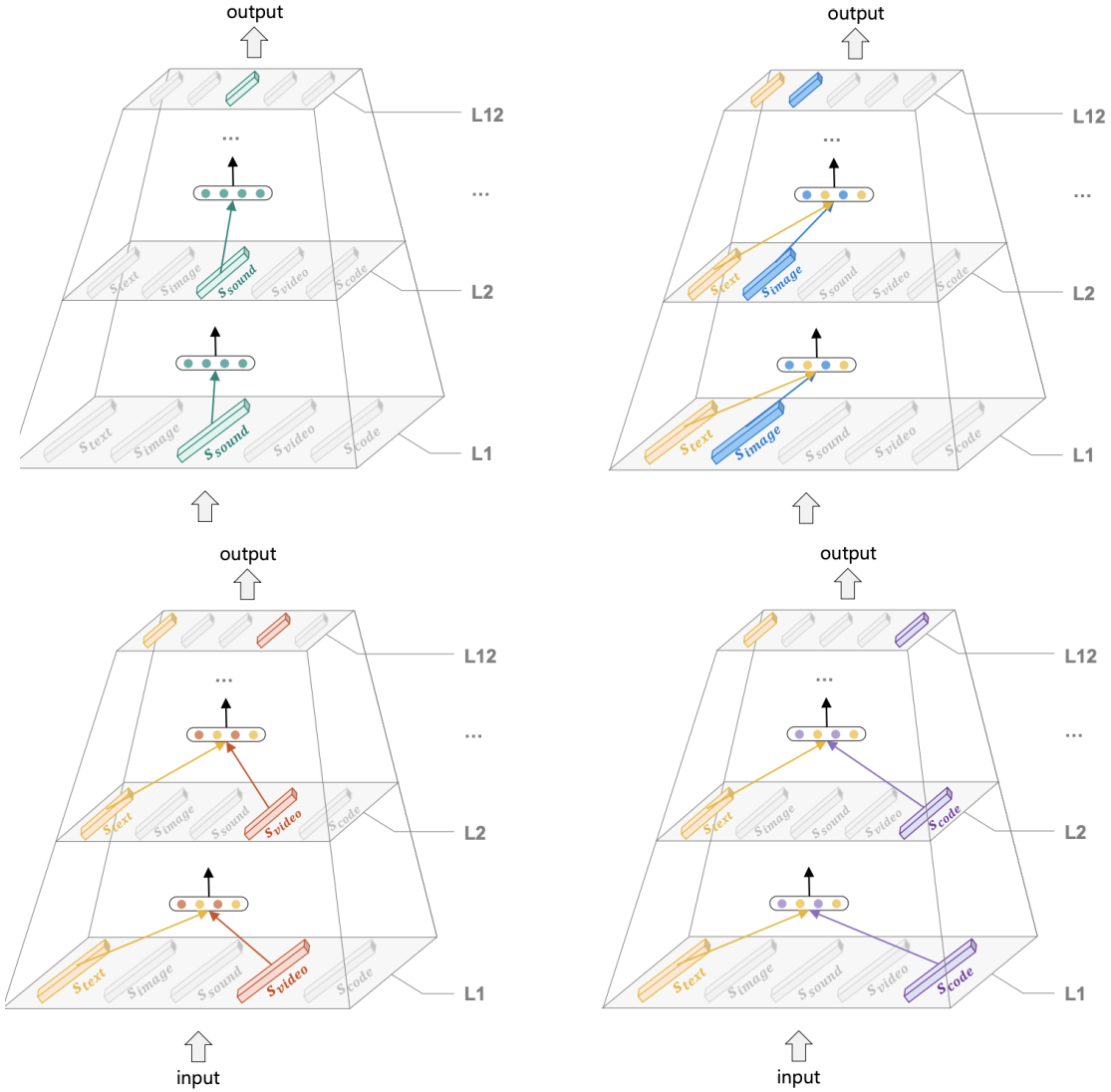
Sparse multimodal Mixture-of-Expert (MoE)

- Handling multiple modalities of information with a single model
- Different parts of the parameters are specialized for processing different modalities



(a) Fully activated dense model

(b) Sparsely activated MoE

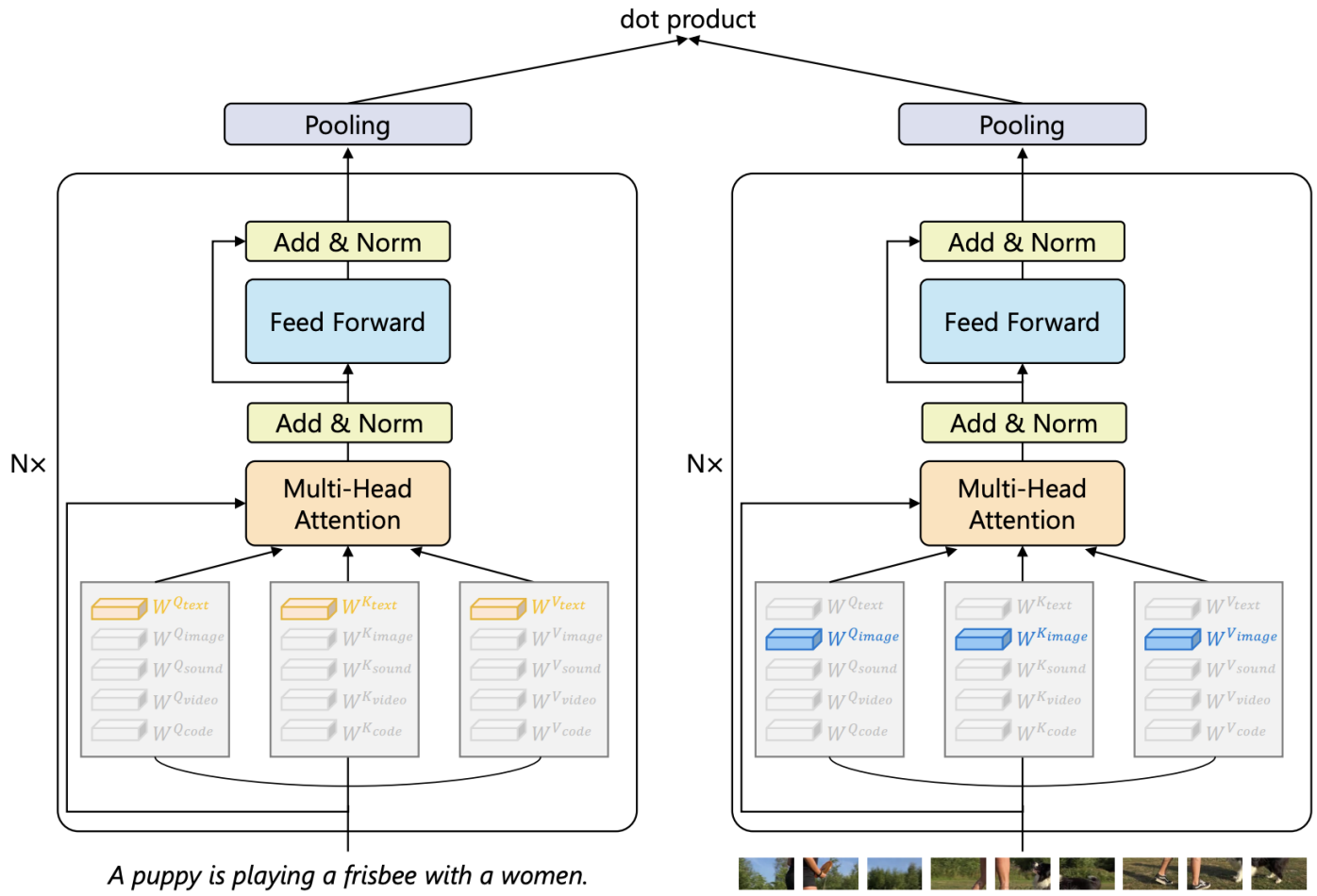


(e) SkillNet for text-to-video retrieval

(f) SkillNet for text-to-code retrieval

SkillNet: One Model, Multiple Modalities: A Sparsely Activated Approach for Text, Sound, Image, Video and Code, arXiv 2022

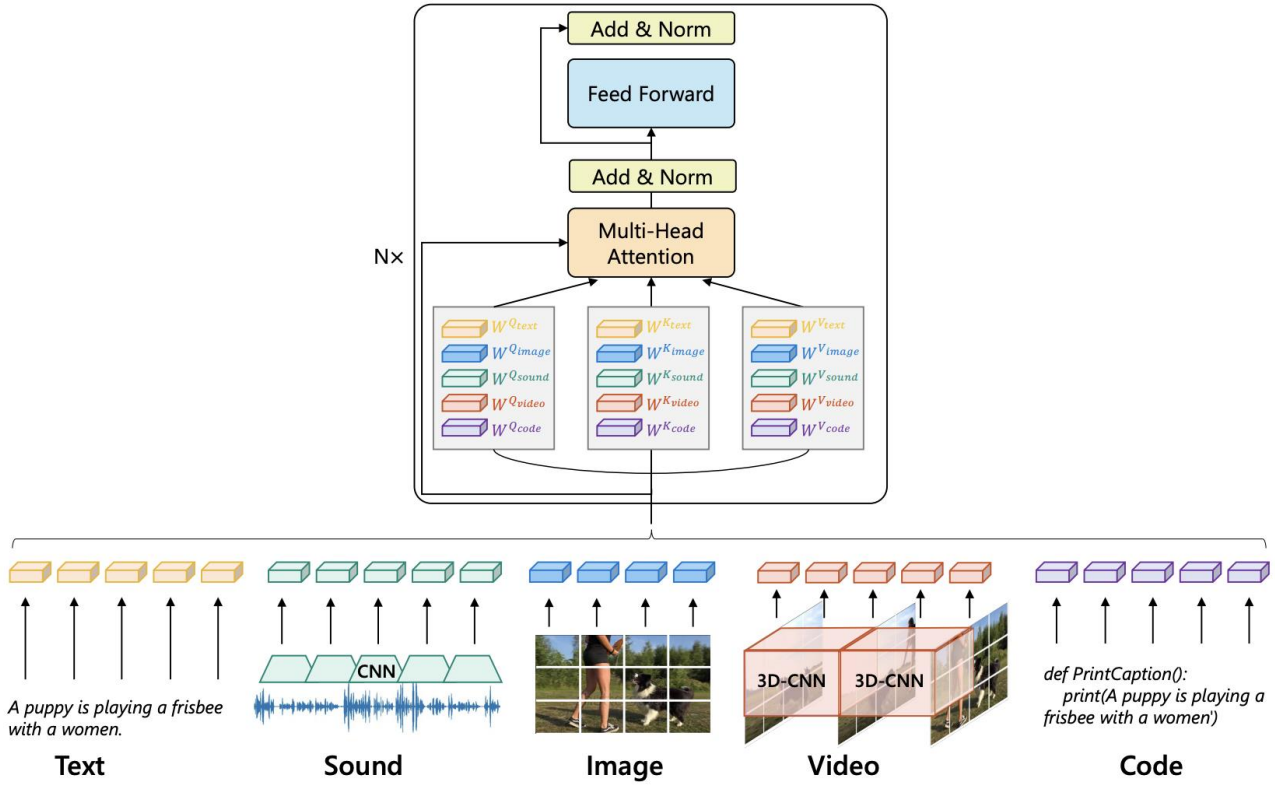
Sparse multimodal Mixture-of-Expert (MoE)



SkillNet: One Model, Multiple Modalities: A Sparsely Activated Approach for Text, Sound, Image, Video and Code, arXiv 2022

Sparse multimodal Mixture-of-Expert (MoE)

- The embeddings of different modalities
  - Text
  - Sound
  - Image
  - Video
  - Code

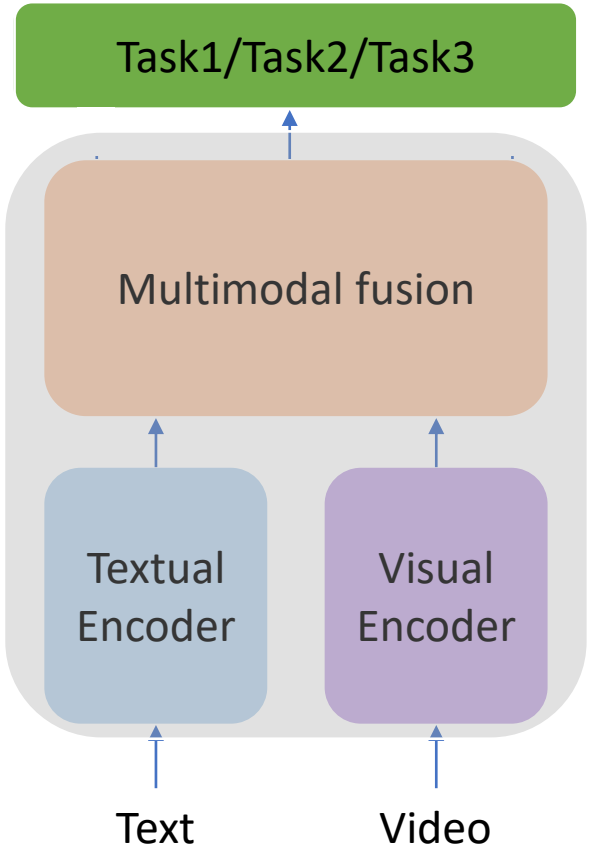


SkillNet: One Model, Multiple Modalities: A Sparsely Activated Approach for Text, Sound, Image, Video and Code, arXiv 2022

Unification

- Modalities: architecture (across audio, code)
  - Uni-Perceiver, All in One, SkillNet

- Tasks: unify VL tasks/V-only tasks as text generation
  - LAVENDER, Socratic Models, GIT, Flamingo



Unify VidL tasks as text generation

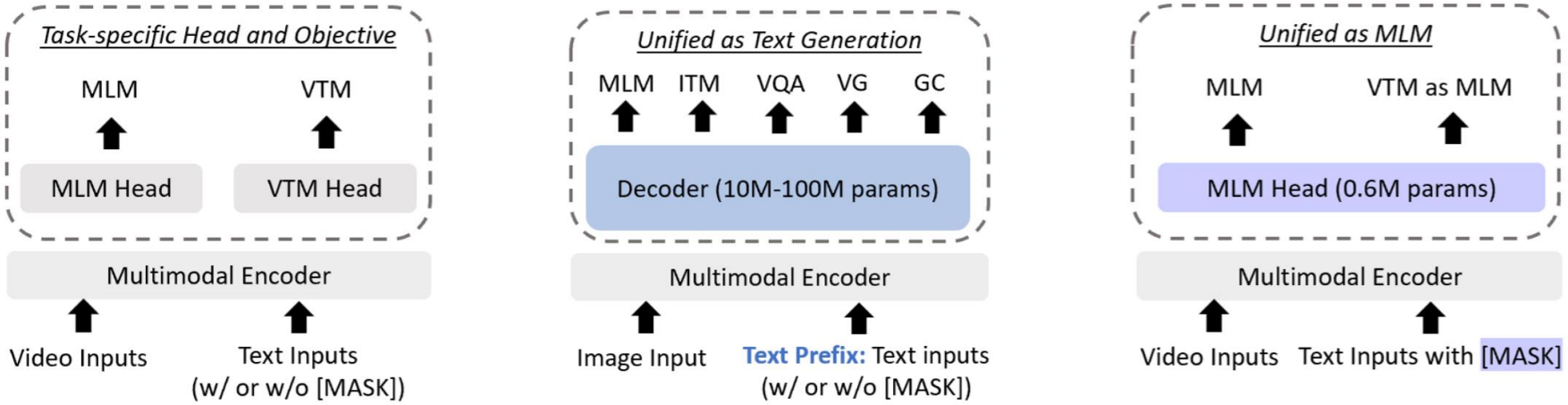
- Unify both pre-training and downstream finetuning as Masked Language Modeling

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling, arXiv 2022

Unify VidL tasks as text generation

- Adopt an encoder-only architecture, with a lightweight MLM head, instead of the heavy decoder in unified image-text models

Comparison to existing methods during pre-training



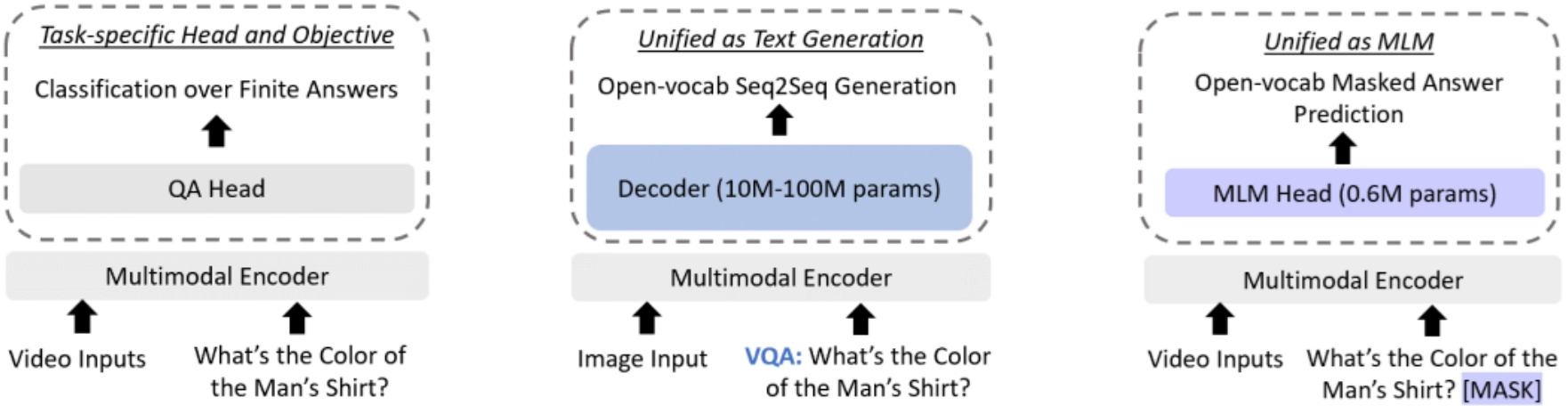
(a) Existing VidL Methods (b) Existing Unified Image-text Models (c) LAVENDER

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling, arXiv 2022

Unify VidL tasks as text generation

- Adopt an encoder-only architecture, with a lightweight MLM head, instead of the heavy decoder in unified image-text models

Comparison to existing methods on downstream image/video question answering task



(a) Existing VidL Methods (b) Existing Unified Image-text Models (c) LAVENDER

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling, arXiv 2022

Unify VidL tasks as text generation

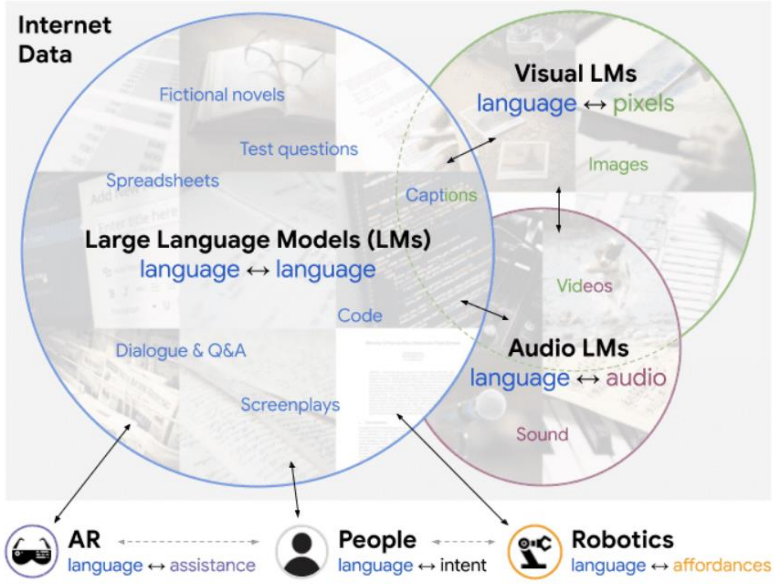
- LAVENDER surpasses previous methods with task-specific designs on 12 video-language tasks

	TGIF			MSRVTT		LSMDC		MSVD	Captioning		Retrieval	
	Action	Transition	Frame	MC	QA	MC	FiB	QA	MSRVTT	MSVD	DiDeMo	LSMDC
Published SOTA	94.0 [84]	96.2 [84]	69.5 [84]	90.9 [84]	43.1 [84]	81.7 [84]	52.9 [84]	46.3 [78]	60.0 [57]	120.6 [35]	65.1 [6]	41.9 [12]
LAVENDER	96.6	99.1	73.5	97.4	45.0	87.0	57.1	56.6	60.1	150.7	72.4	43.3
Δ	2.6↑	2.9↑	4.0↑	6.5↑	1.9↑	5.3↑	4.2↑	10.3↑	0.1↑	30.1↑	7.3↑	1.4↑

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling, arXiv 2022

Composing Zero-Shot Multimodal Reasoning with Language

- Different pretrained models store different forms of commonsense knowledge across different domains
  - Visual-language models (VLMs) are trained on Internet-scale image captions
  - Large language models (LMs) are further trained on Internet-scale text with no images (e.g., spreadsheets, SAT questions, code)
- The diversity is symbiotic




```
I am an intelligent image captioning bot. This image is a {img_type}. There {num_people}. I think this photo was taken at a {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, {object3},... in this {img_type}. A creative short caption I can generate to describe this image is:
```

SM (ours): This image shows an inviting dining space with plenty of natural light.

ClipCap: A wooden table sitting in front of a window.

SM (ours): People gather under a blossoming cherry tree, enjoying the beauty of nature together.

ClipCap: Students enjoying the cherry blossoms.

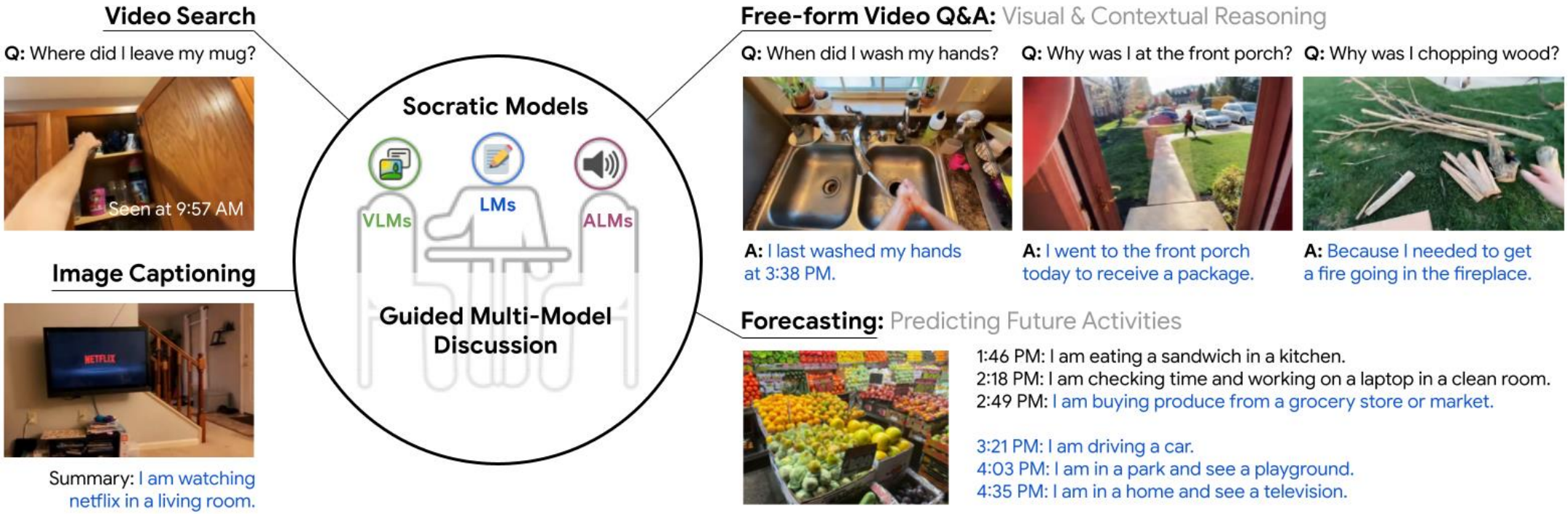
SM (ours): At the outdoor market, you can find everything from plantains to Japanese bananas.

ClipCap: A bunch of bananas sitting on top of a table.

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, arXiv 2022

Composing Zero-Shot Multimodal Reasoning with Language

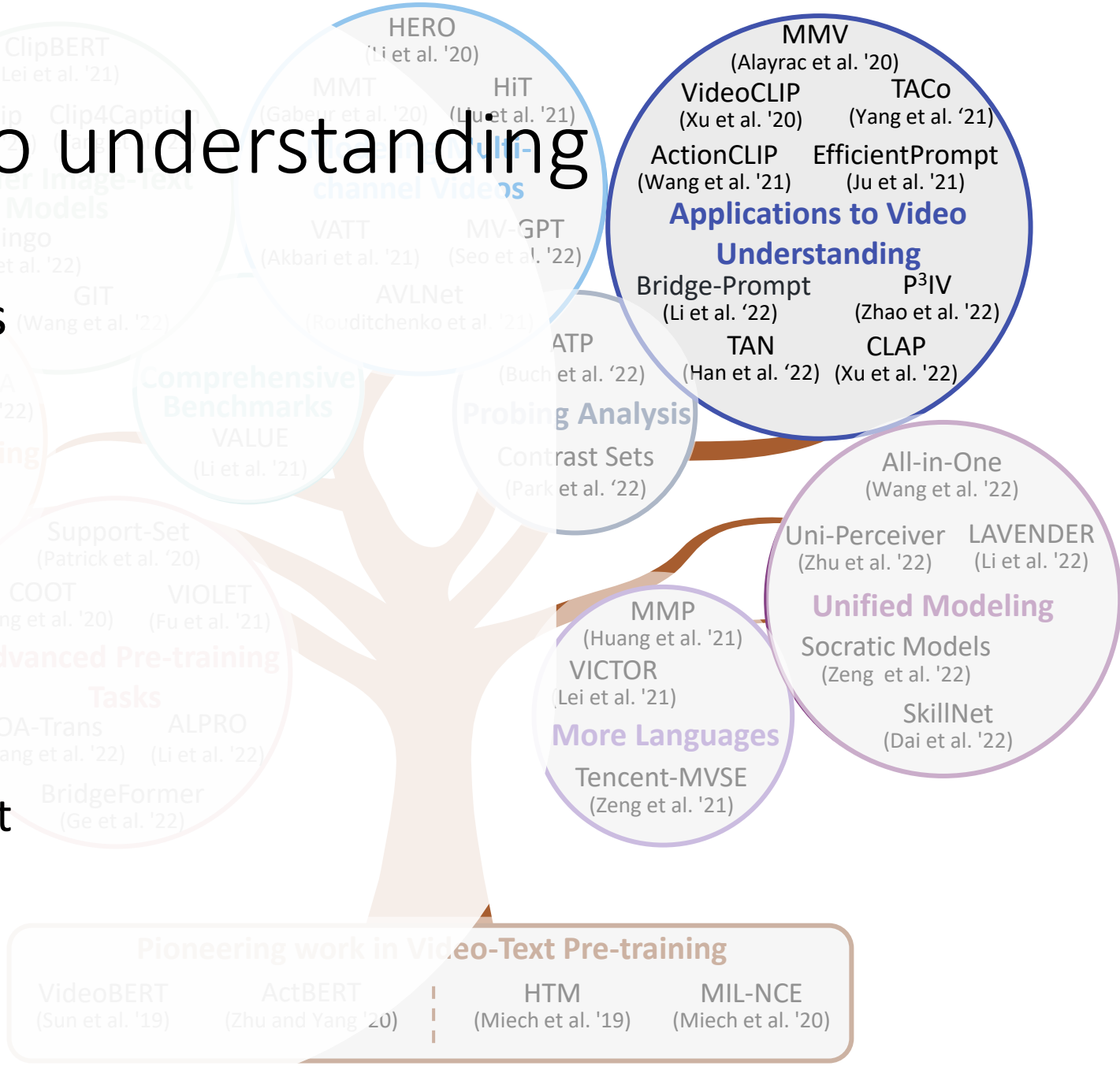
- A modular framework in which multiple pretrained models may be composed zero-shot
  - via multimodal-informed prompting, to exchange information with each other and capture new multimodal capabilities, without requiring finetuning.



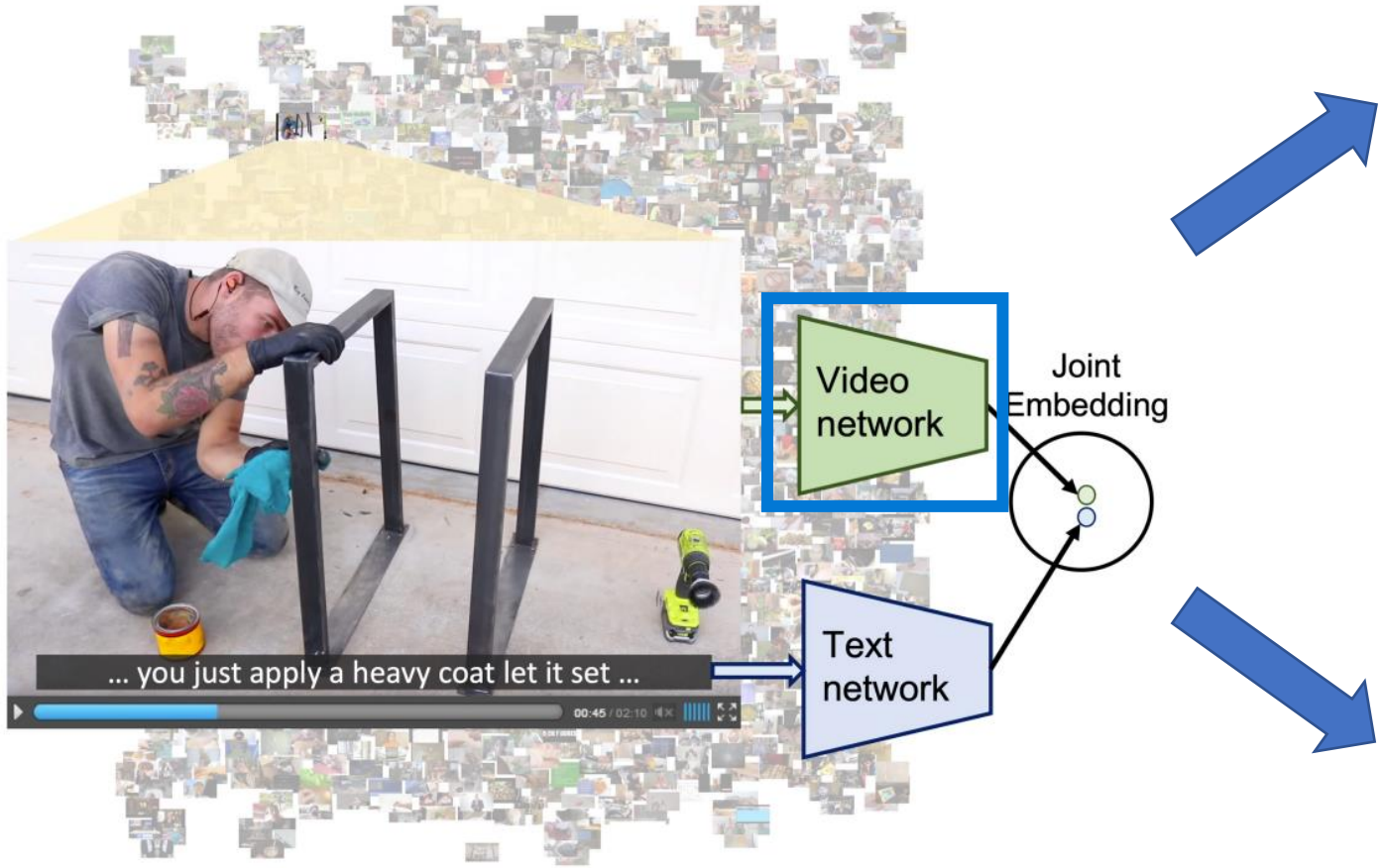
Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, arXiv 2022

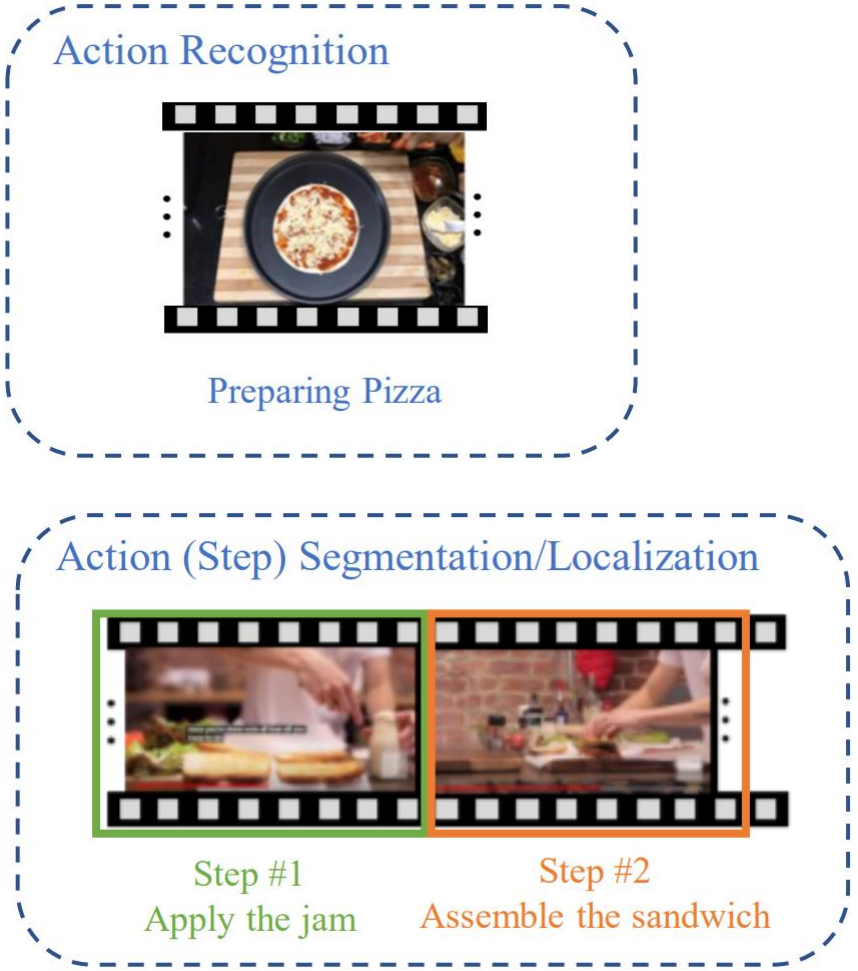
Applications to video understanding

- Video-text contrastive models
  - MMV, VideoCLIP, TACo
- Transfer image-text models
  - ActionCLIP
- With Prompt Tuning
  - EfficientPrompt, Bridge-Prompt



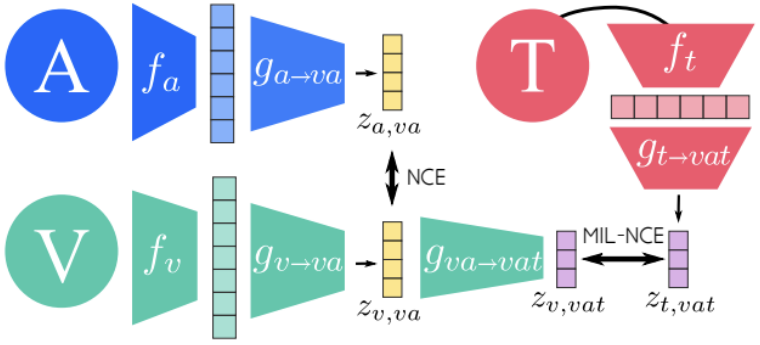
Video-text contrastive models



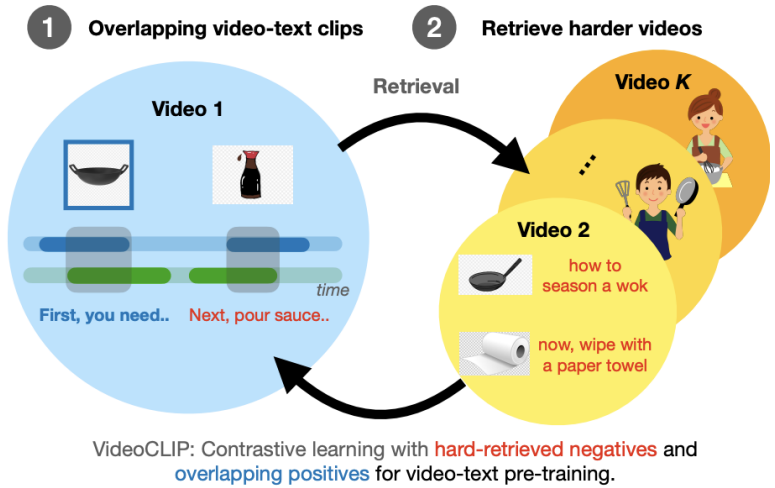


[HTM, Miech et al., 2019], [MIL-NCE, Miech et al., 2020]
Figure credit: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, ICCV 2019

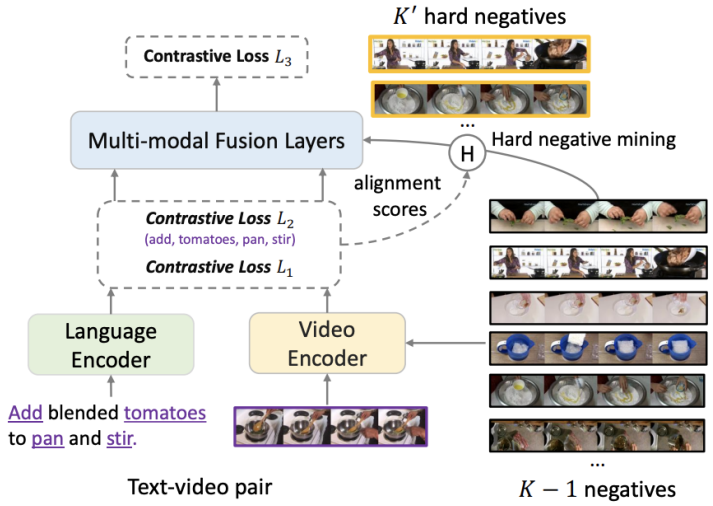
Video-text contrastive models



MMV: Self-Supervised MultiModal Versatile Networks, NeurIPS 2020

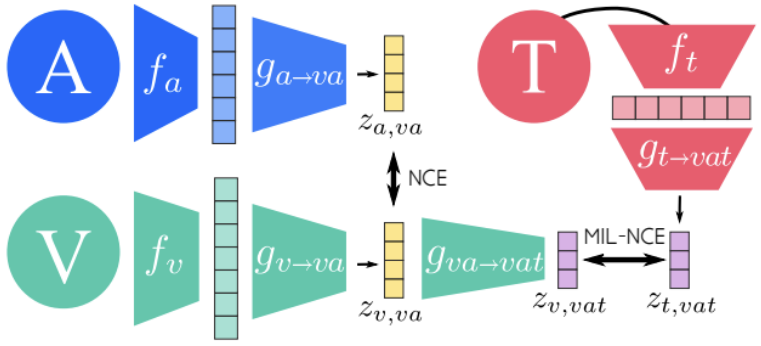


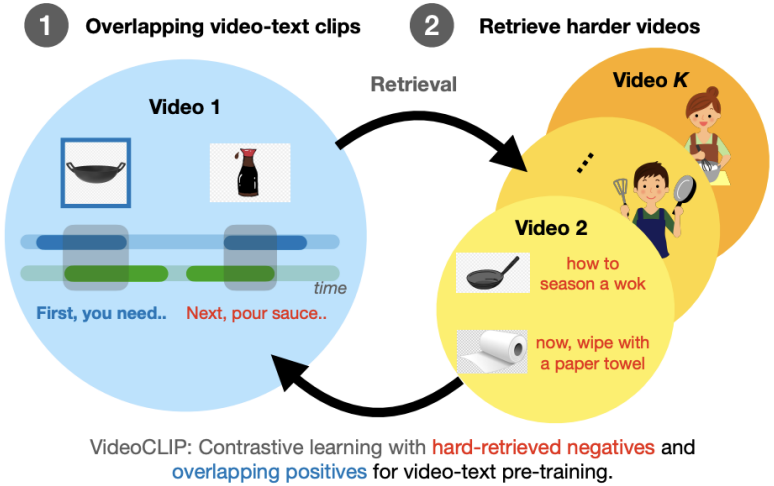
VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021

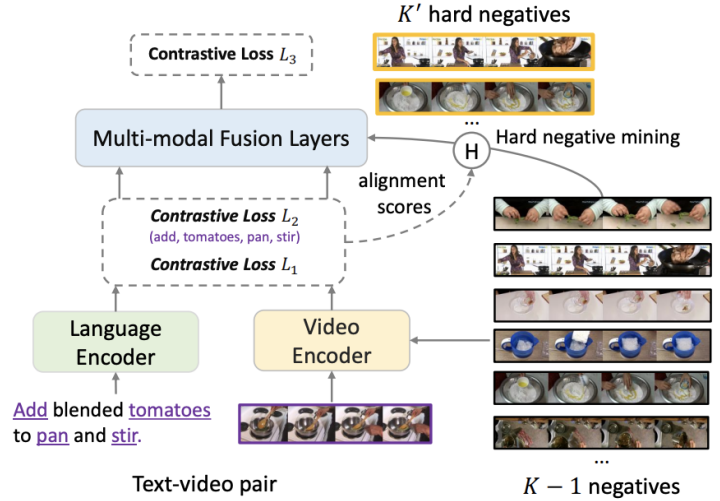


TACo: Token-aware Cascade Contrastive Learning for Video-Text Alignment, ICCV 2021

Video-text contrastive models







MMV: Self-Supervised MultiModal Versatile Networks, NeurIPS 2020

VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021

TACo: Token-aware Cascade Contrastive Learning for Video-Text Alignment, ICCV 2021

$$\mathcal{L}(x) = \lambda_{va}\text{NCE}(x_v, x_a) + \lambda_{vt}\text{MIL-NCE}(x_v, x_t)$$

$$\mathcal{L} = -\sum_{(v,t)\in B}\left(\log \text{NCE}(z_v, z_t) + \log \text{NCE}(z_t, z_v)\right)$$

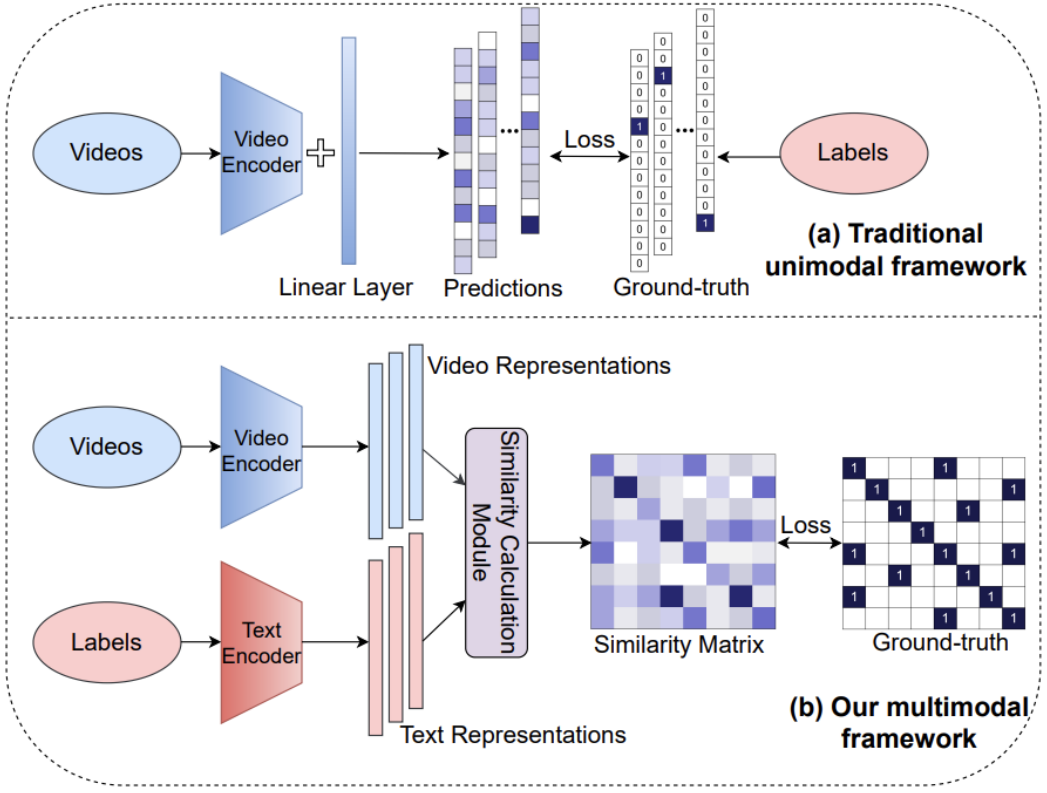
$$L_1 = -\sum_{i=1}^{K}\log\left(\frac{\exp^{\bar{x}_i\cdot\bar{y}_i/\tau_1}}{\exp^{\bar{x}_i\cdot\bar{y}_i/\tau_1} + \sum_{j\neq i}\exp^{\bar{x}_j\cdot\bar{y}_i/\tau_1}}\right)$$

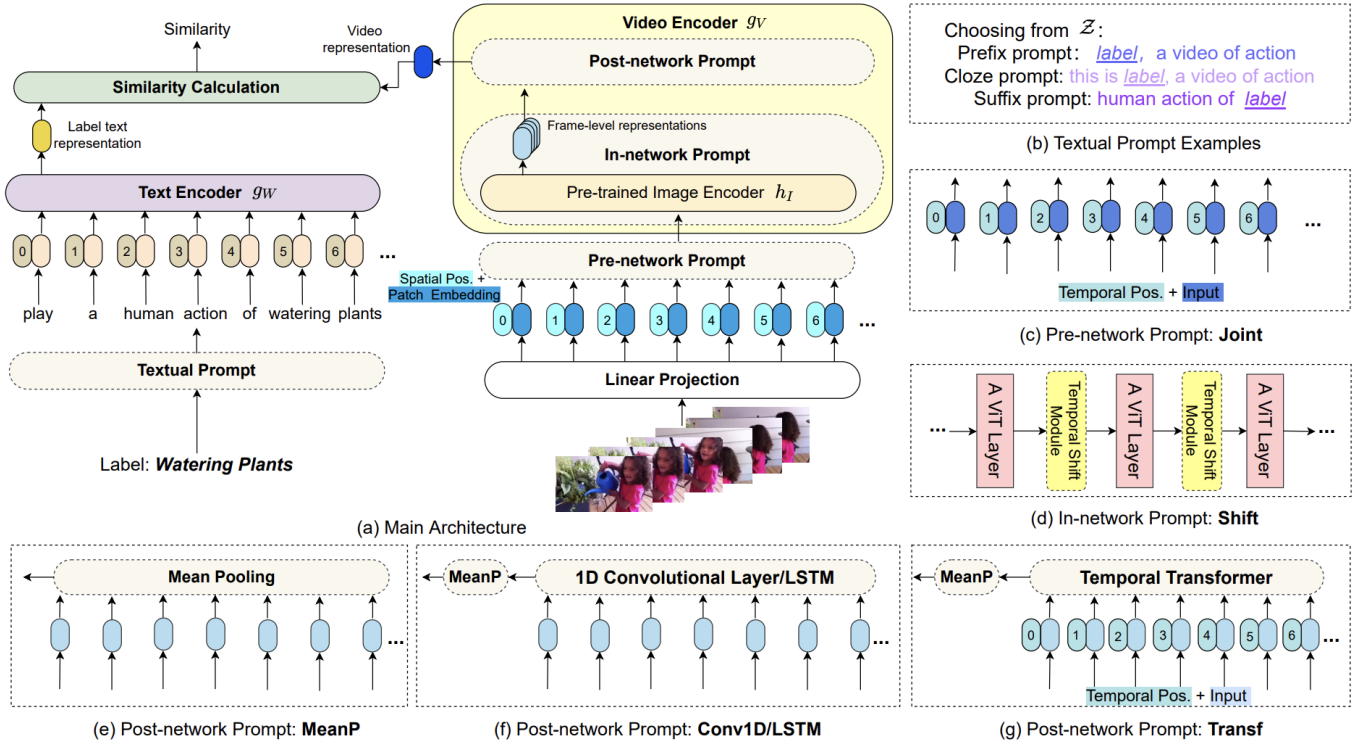
$$L_2 = -\sum_{i=1}^{K}\sum_{p\in\mathcal{P}_i}\log\left(\frac{\exp^{s(x_i,y_i^p)/\tau_2}}{\exp^{s(x_i,y_i^p)/\tau_2} + \sum_{j\neq i}\exp^{s(x_j,y_i^p)/\tau_2}}\right)$$

$$L_3 = -\sum_{i=1}^{K}\log\left(\frac{\exp^{w\cdot z_{i,i}^{cls}}}{\exp^{w\cdot z_{i,i}^{cls}} + \sum_{j\neq i}\exp^{w\cdot z_{j,i}^{cls}}}\right)$$

Transfer image-text models

- Attaching importance to the semantic information of label texts rather than simply mapping them into numbers.

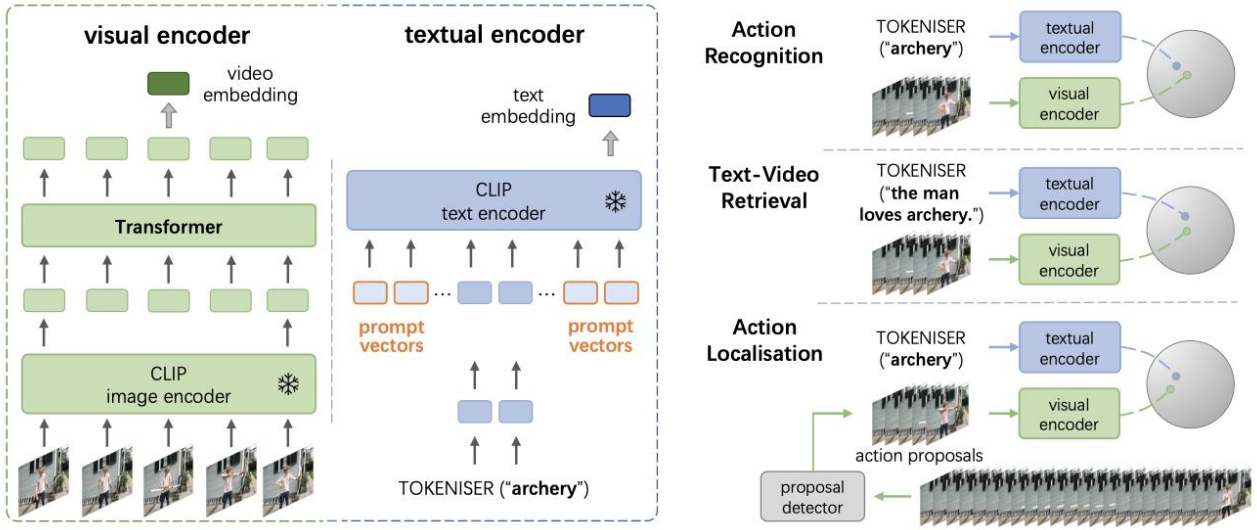




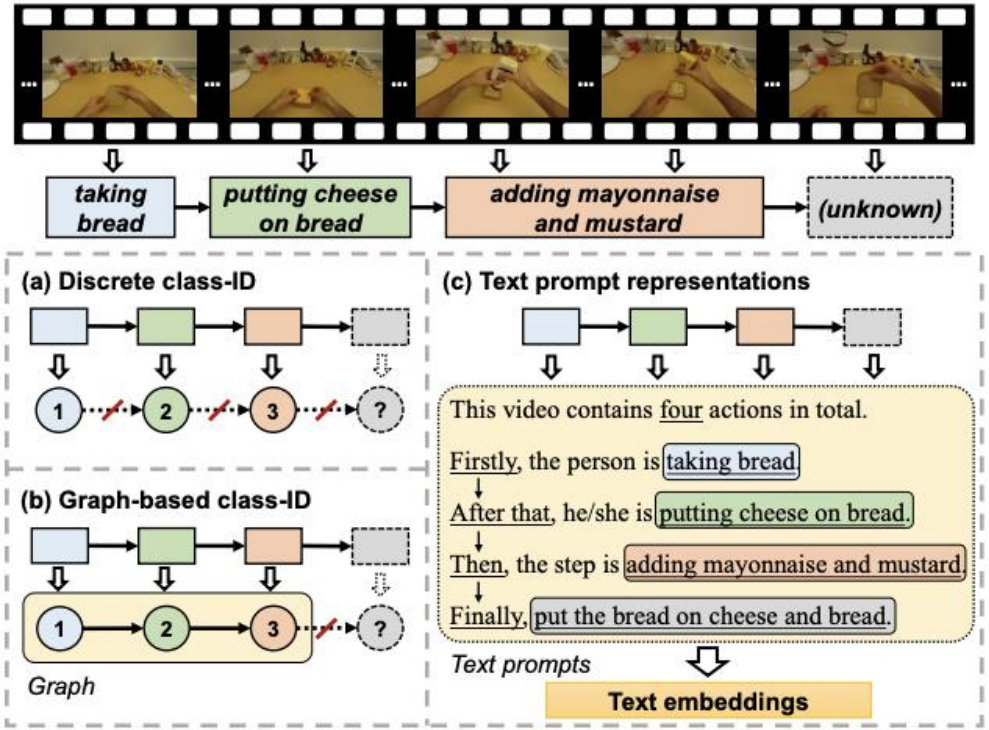
ActionCLIP: A New Paradigm for Video Action Recognition, arXiv 2021
CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval, arXiv 2021

Prompting

- Efficiently adapt one pre-trained visual-language model to novel tasks with minimal training



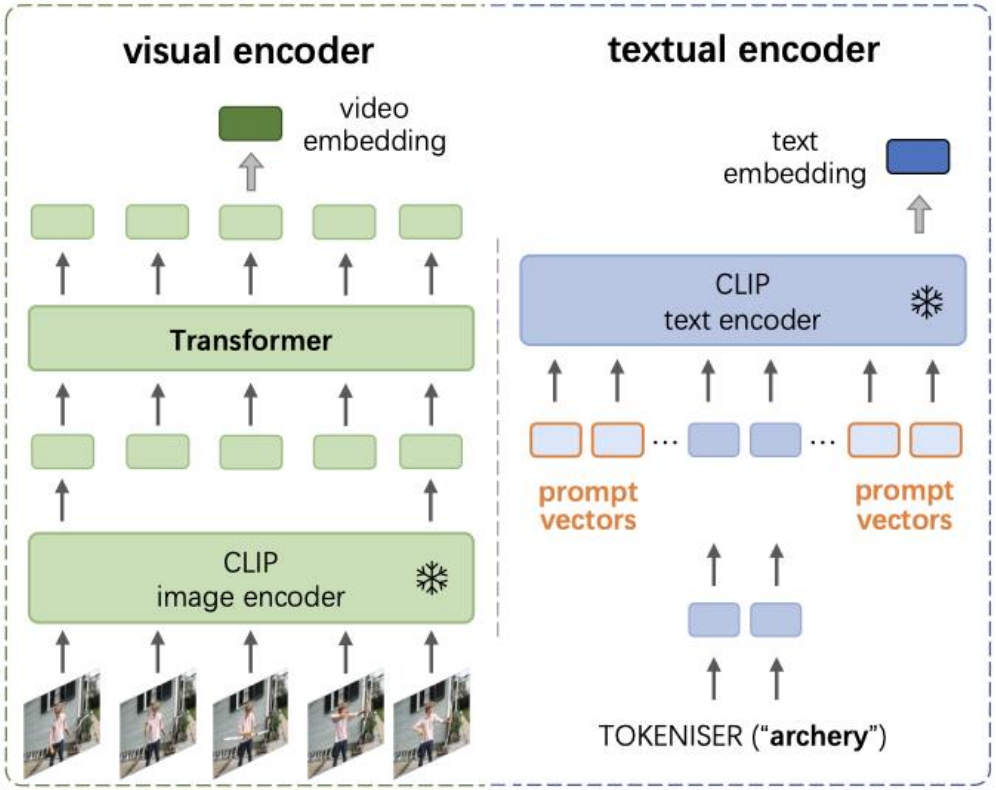
Efficient Prompting: Prompting Visual-Language Models for Efficient Video Understanding, arXiv 2021

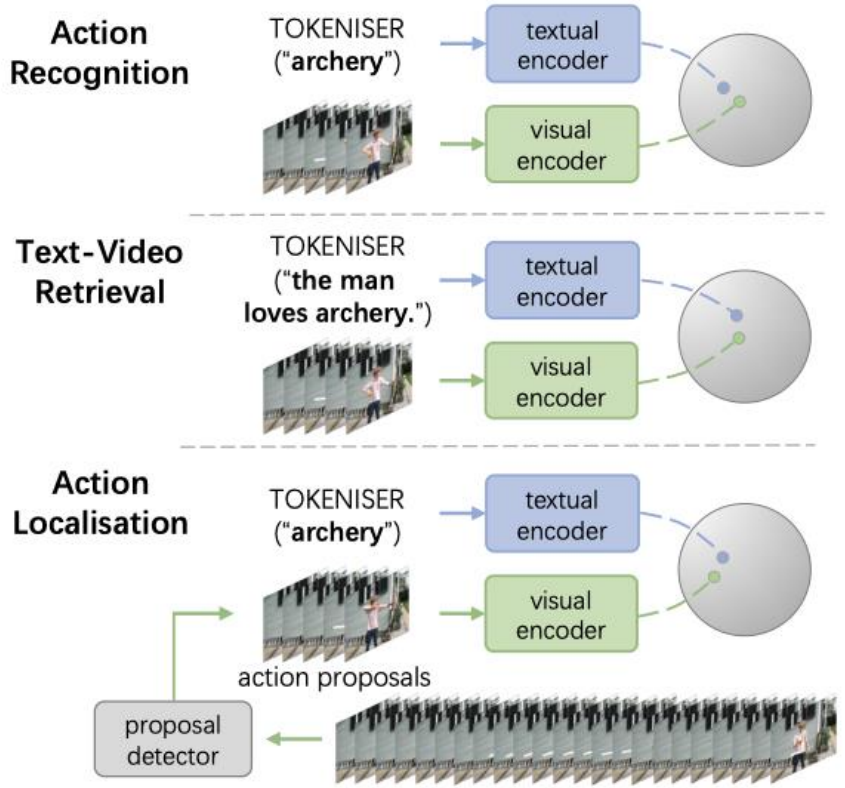


Bridge-Prompt: Towards Ordinal Action Understanding in Instructional Videos, CVPR 2022

Continuous prompt vectors

- Optimize a few random vectors that convert the novel tasks into the same format as the pre-training objectives

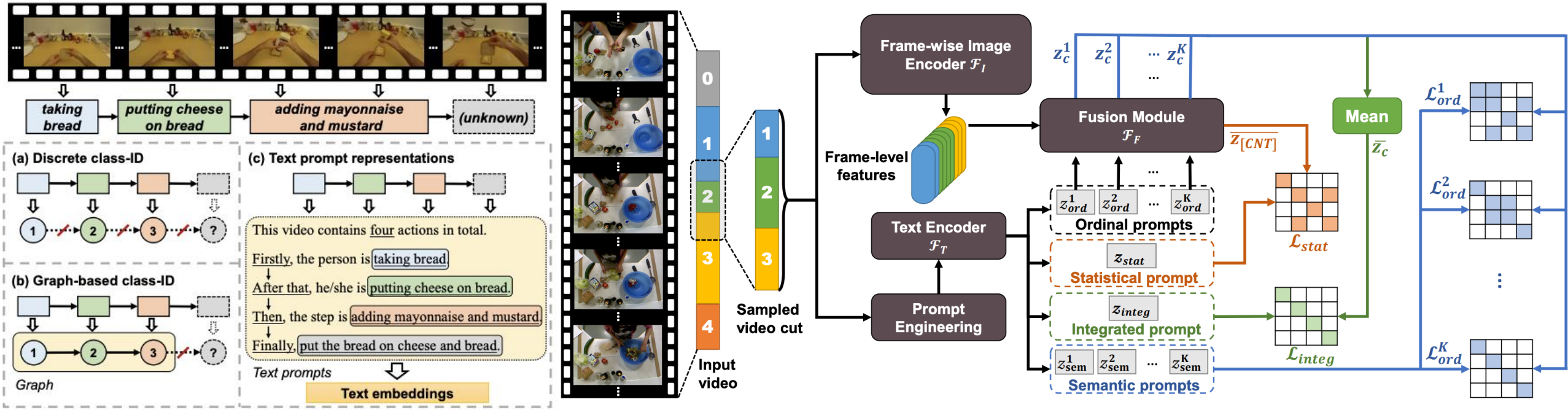




Efficient Prompting: Prompting Visual-Language Models for Efficient Video Understanding, arXiv 2021

Model the semantics across adjacent actions

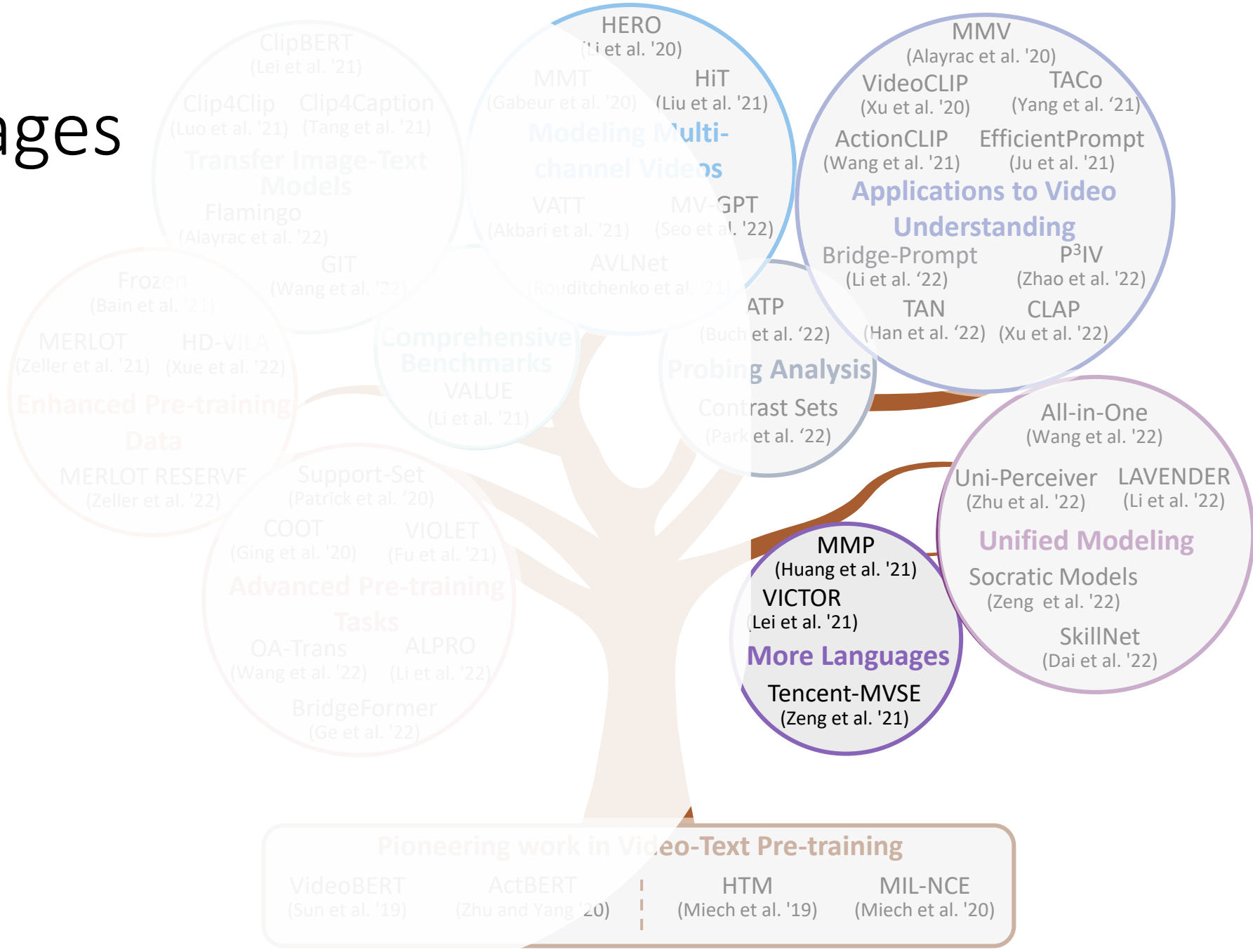
- Simultaneously exploits both out-of-context and contextual information from a series of ordinal actions in instructional videos



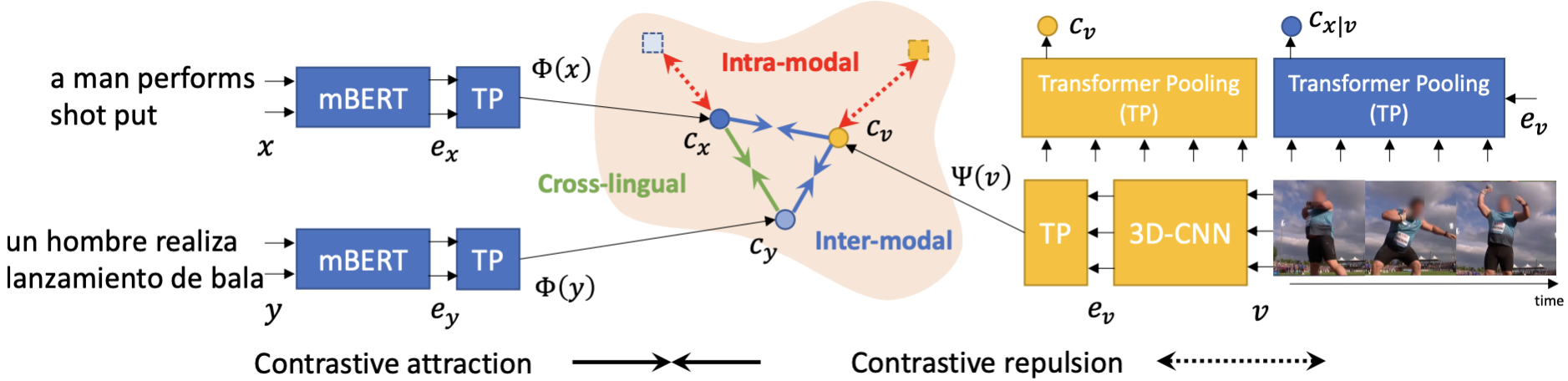
Bridge-Prompt: Towards Ordinal Action Understanding in Instructional Videos, CVPR 2022

More languages

- MMP
- VICTOR
- Tencent-MVSE

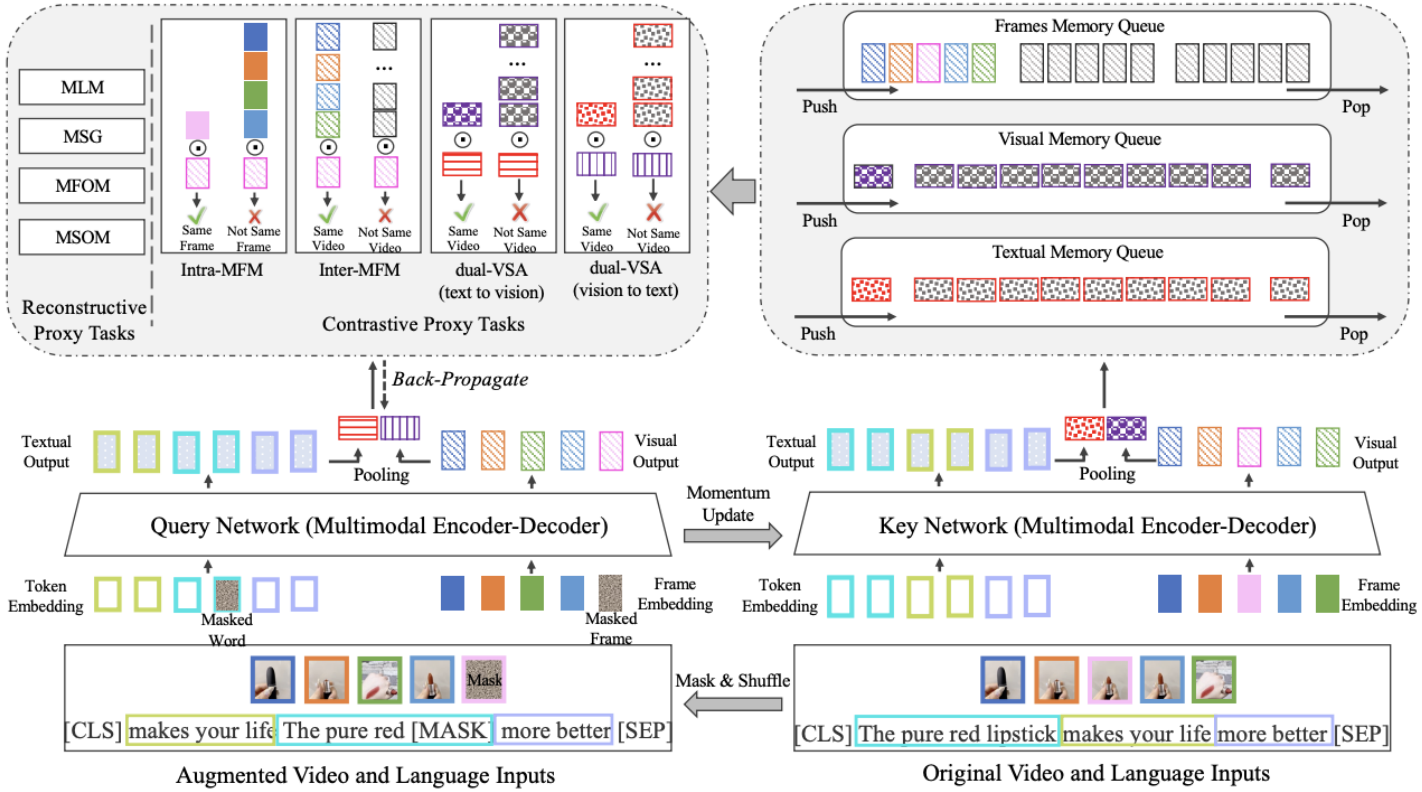


Multilingual Multimodal Pre-training



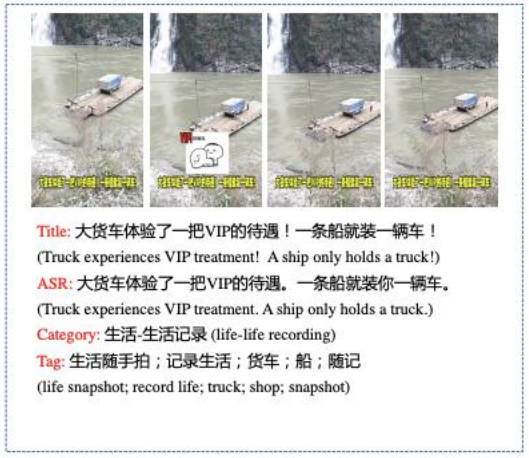
MMP: Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models, NAACL 2021

Understanding Chinese Video and Language via Contrastive Multimodal Pre-Training



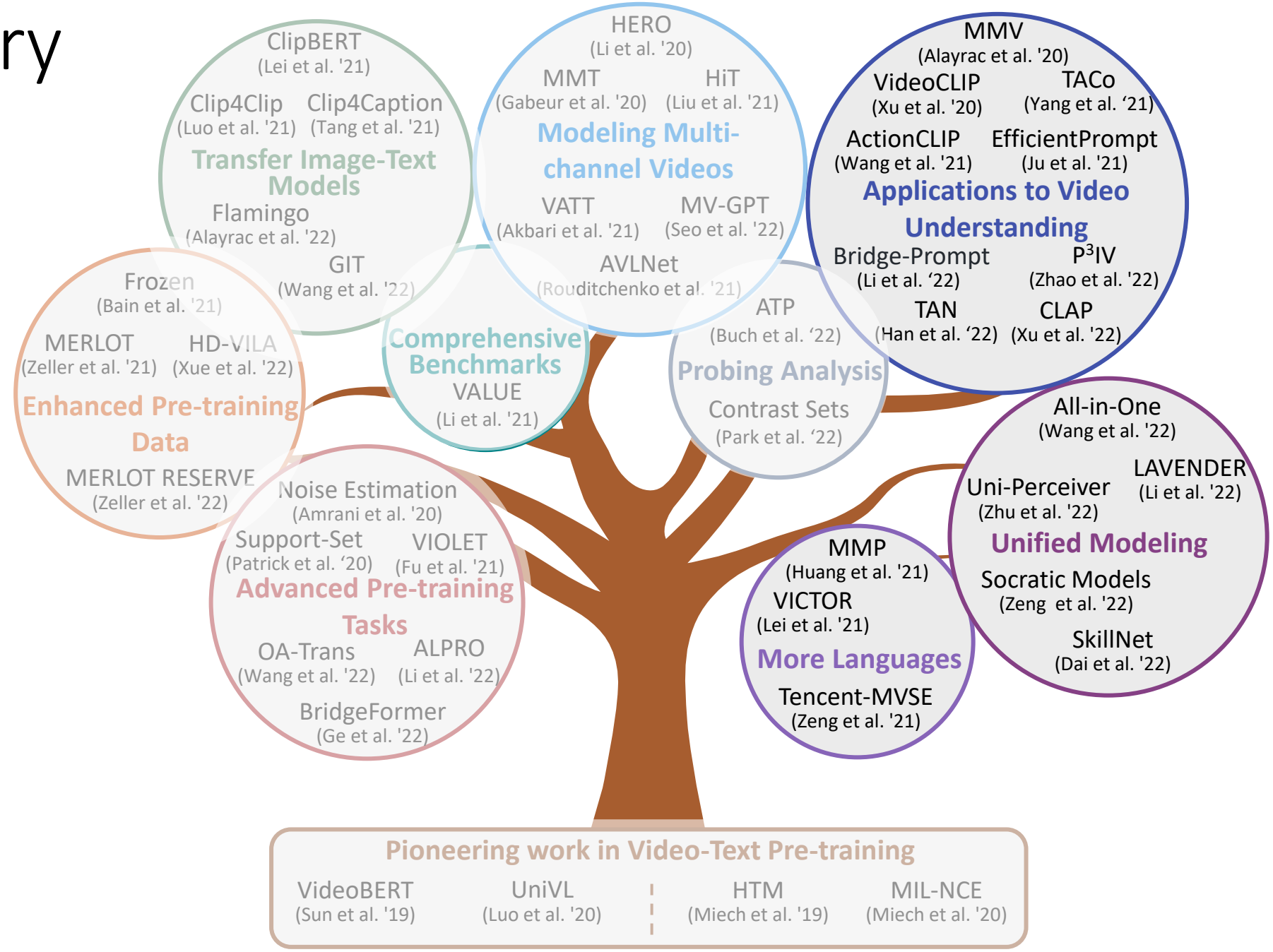
VICTOR: Understanding Chinese Video and Language via Contrastive Multimodal Pre-Training, arXiv 2021

A Large-Scale Benchmark Dataset for Multi-Modal Video Similarity Evaluation



Tencent-MVSE: A Large-Scale Benchmark Dataset for Multi-Modal Video Similarity Evaluation, CVPR 22

Summary


ClipBERT
(Lei et al. '21)
Clip4Clip
(Luo et al. '21)
Clip4Caption
(Tang et al. '21)
Transfer Image-Text Models
Flamingo
(Alayrac et al. '22)
GIT
(Wang et al. '22)
HERO
(Li et al. '20)
MMT
(Gabeur et al. '20)
HiT
(Liu et al. '21)
Modeling Multi-channel Videos
VATT
(Akbari et al. '21)
MV-GPT
(Seo et al. '22)
AVLNet
(Rouditchenko et al. '21)
MMV
(Alayrac et al. '20)
VideoCLIP
(Xu et al. '20)
TACo
(Yang et al. '21)
ActionCLIP
(Wang et al. '21)
EfficientPrompt
(Ju et al. '21)
Applications to Video Understanding
Bridge-Prompt
(Li et al. '22)
P3IV
(Zhao et al. '22)
TAN
(Han et al. '22)
CLAP
(Xu et al. '22)
Frozen
(Bain et al. '21)
MERLOT
(Zeller et al. '21)
HD-VILA
(Xue et al. '22)
Enhanced Pre-training Data
MERLOT RESERVE
(Zeller et al. '22)
Comprehensive Benchmarks
VALUE
(Li et al. '21)
ATP
(Buch et al. '22)
Probing Analysis
Contrast Sets
(Park et al. '22)
All-in-One
(Wang et al. '22)
Uni-Perceiver
(Zhu et al. '22)
LAVENDER
(Li et al. '22)
Unified Modeling
Socratic Models
(Zeng et al. '22)
SkillNet
(Dai et al. '22)
Noise Estimation
(Amrani et al. '20)
Support-Set
(Patrick et al. '20)
VIOLET
(Fu et al. '21)
Advanced Pre-training Tasks
OA-Trans
(Wang et al. '22)
ALPRO
(Li et al. '22)
BridgeFormer
(Ge et al. '22)
MMP
(Huang et al. '21)
VICTOR
(Lei et al. '21)
More Languages
Tencent-MVSE
(Zeng et al. '21)
Pioneering work in Video-Text Pre-training
VideoBERT
(Sun et al. '19)
UniVL
(Luo et al. '20)
HTM
(Miech et al. '19)
MIL-NCE
(Miech et al. '20)