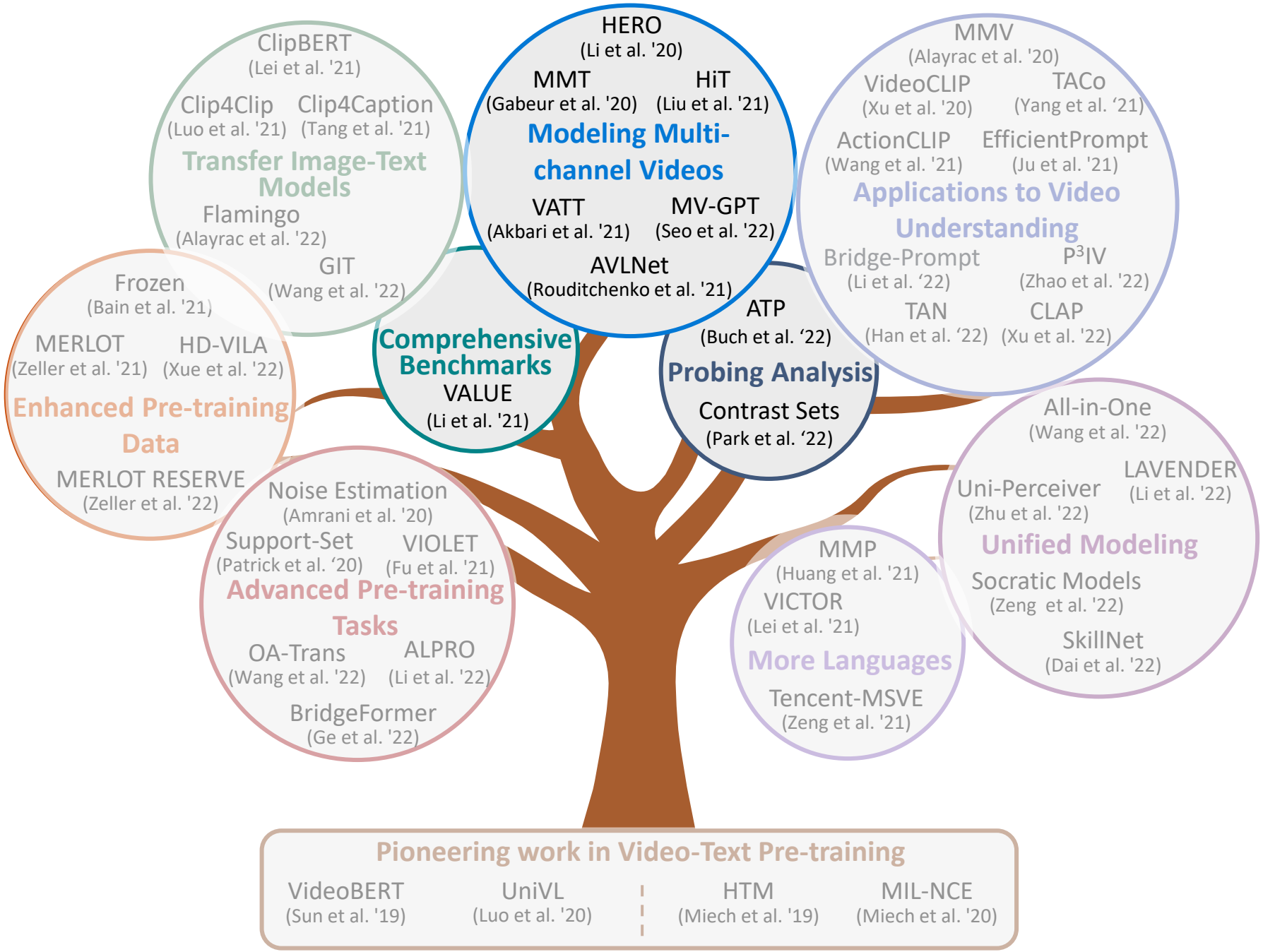
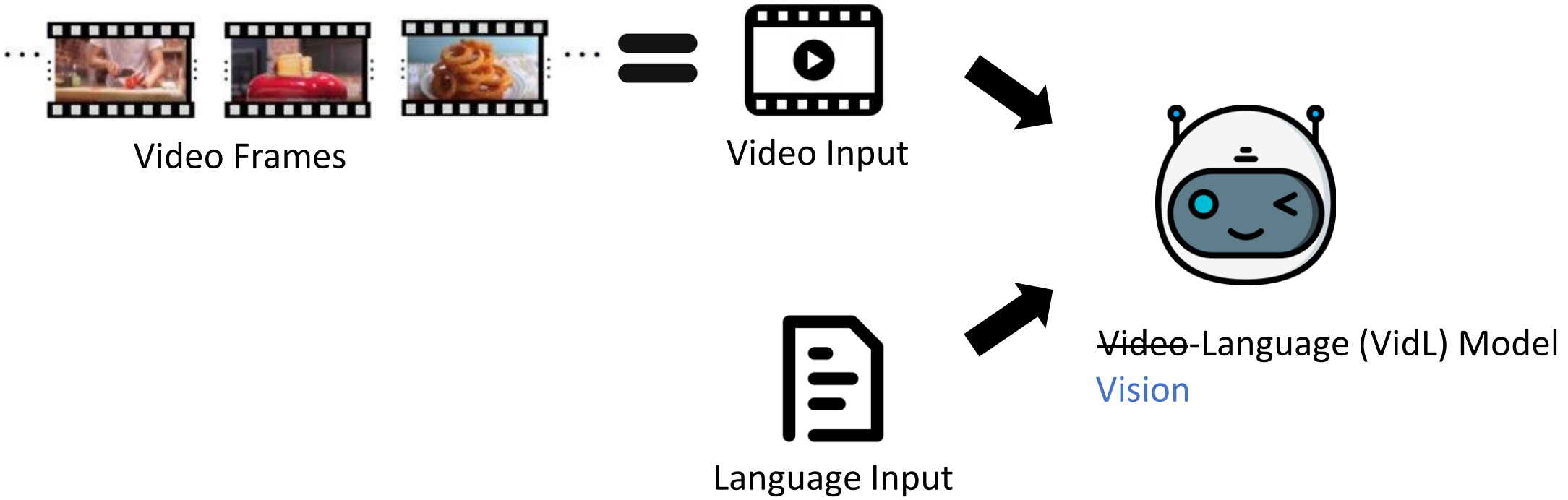
Learning from Multi-channel Videos: Methods and Benchmarks

Linjie Li


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

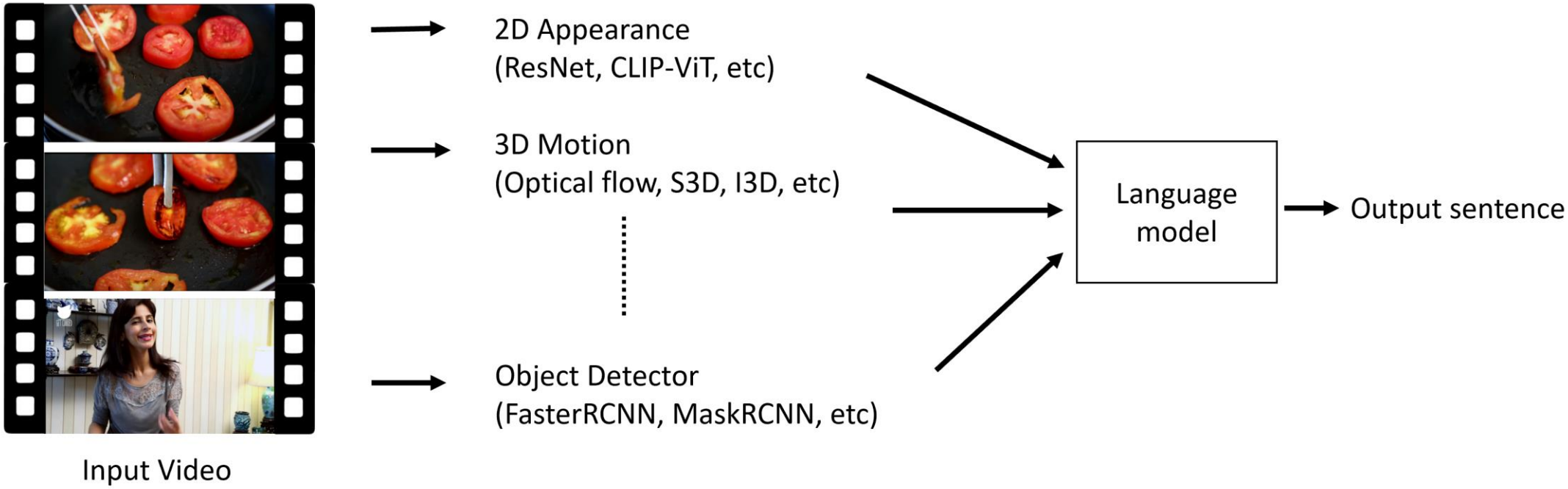
The “V” in Video

Single-channel Video: taking video frames only to represent a video



The “V” in Video

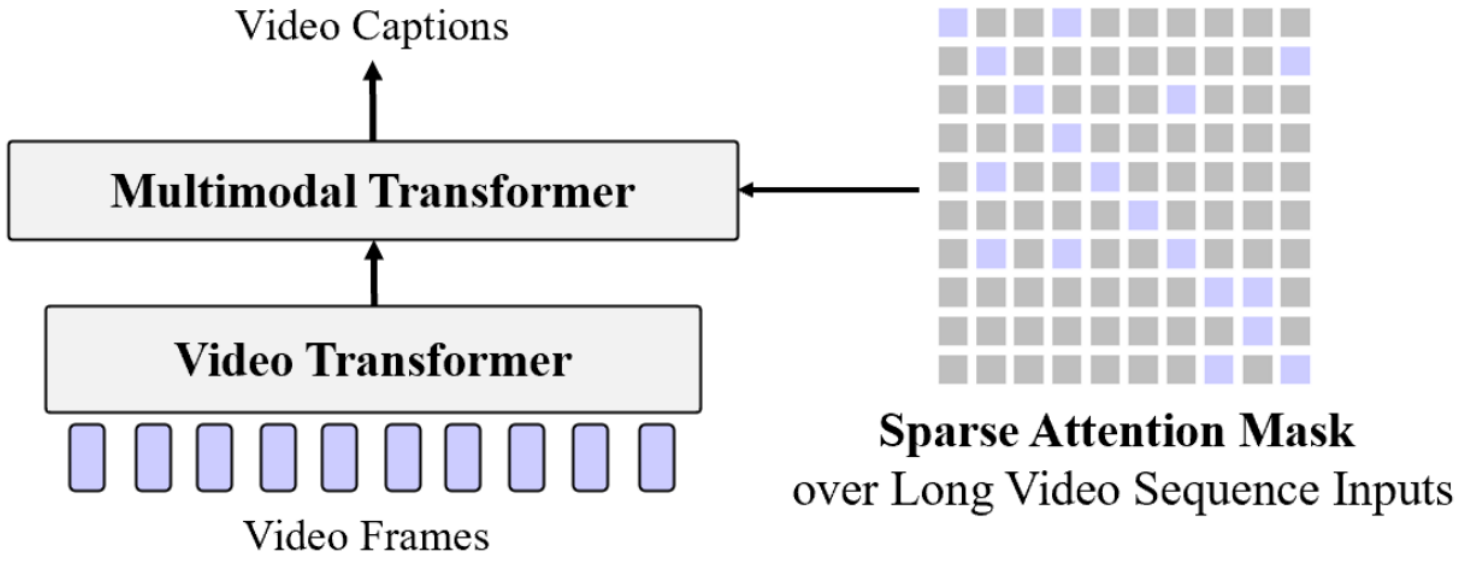
Single-channel Video: taking video frames only to represent a video



Example: leveraging expert vision features to generate captions for single-channel videos

The “V” in Video

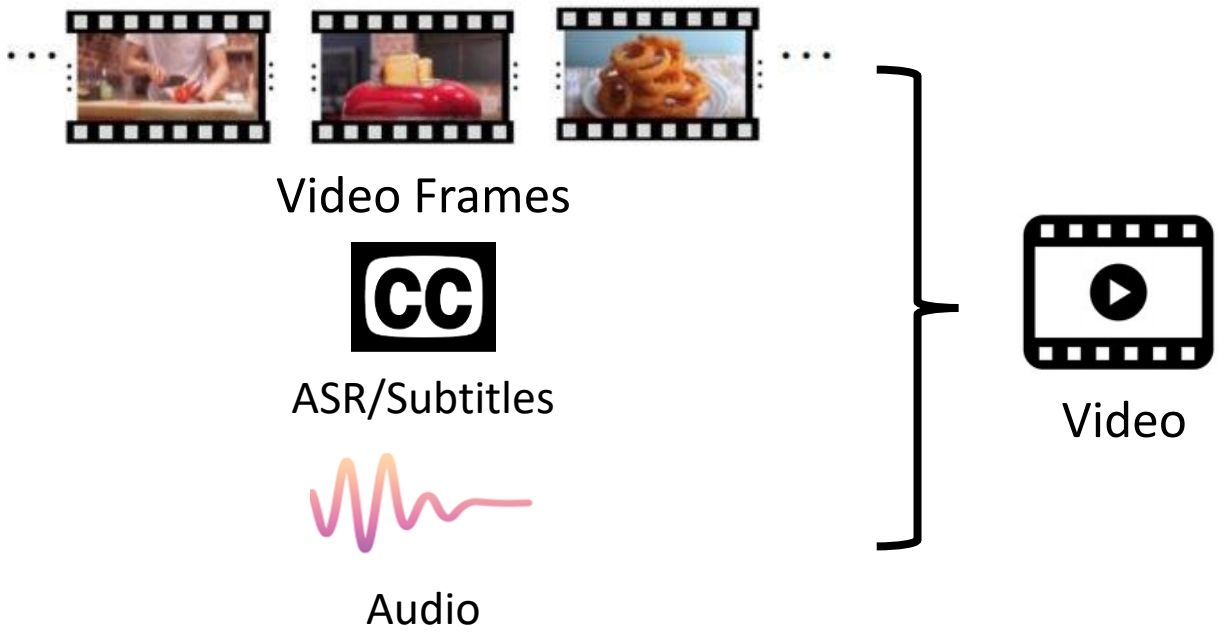
Single-channel Video: taking video frames only to represent a video



SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning, CVPR 2022

Videos are Multi-channel in Nature

Multi-channel Video: Visual Frames + Subtitle + Audio



Videos are Multi-channel in Nature

Examples from Multi-modal Transformer for Video Retrieval, ECCV 2020

Videos are Multi-channel in Nature

Multi-channel Video: Visual Frames + Subtitle + Audio

Video Frames

ASR/Subtitles

Audio

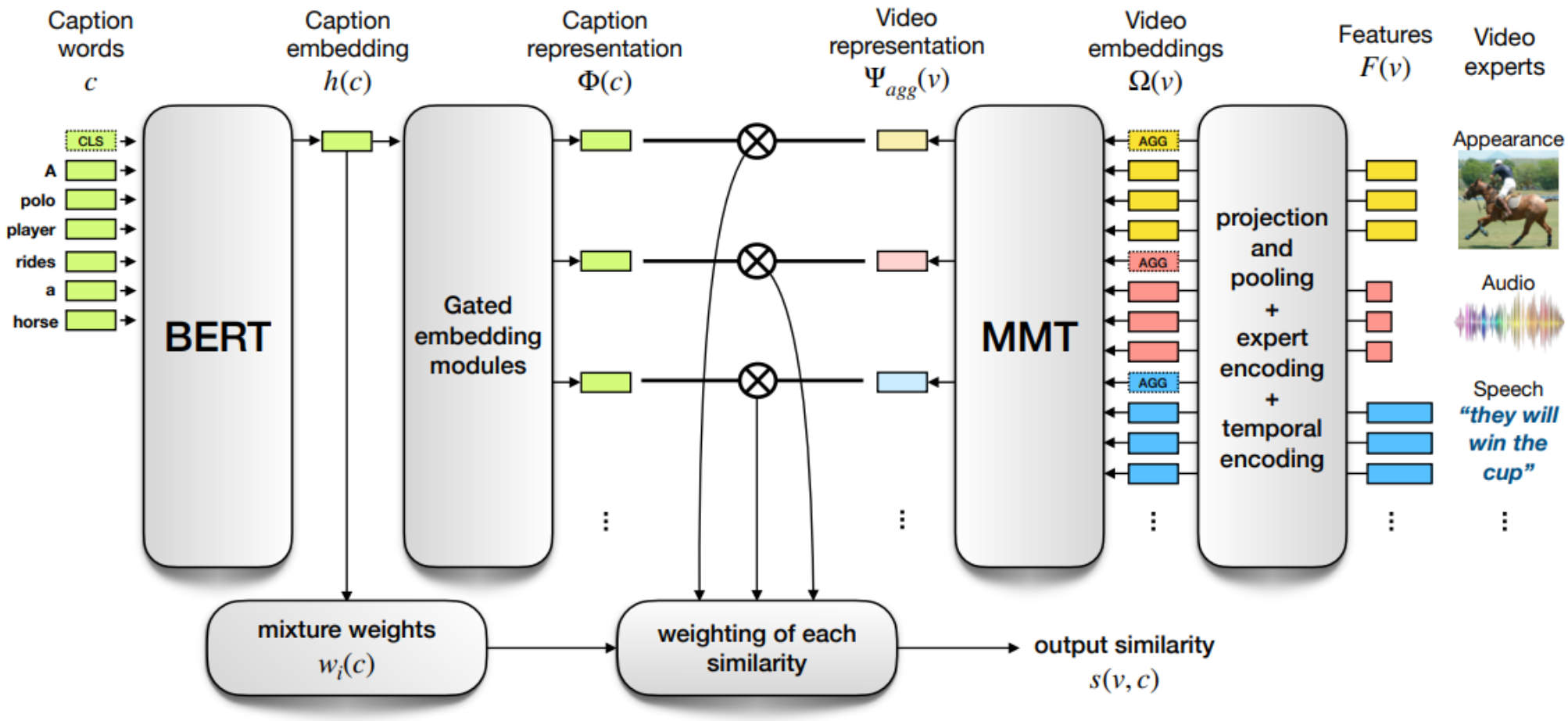
Video Input

Language Input

VidL Model

How to encode the information from all channels?

Modeling Multi-Channel Videos with Expert Features: MMT



Multi-modal Transformer for Video Retrieval, ECCV 2020

Modeling Multi-Channel Videos with Expert Features: MMT

7 Expert Features

- OCR
  - Pre-trained scene text detector -> pre-trained text recognition model trained on Synth90K -> word2vec embeddings
- Speech
  - Speech transcripts extracted using ASR API, embedded with Word2Vec
- Face
  - Pre-trained Face detector -> pre-trained ResNet50 on VGGFace2 for face classification
- Audio
  - Pretrained CNN models for audio recognition on YT8M
- Scene
  - Pre-trained 2D CNN on Place365 for Scene Classification
- Appearance
  - Pre-trained 2D CNN on ImageNet for Image Classification
- Motion
  - Pre-trained 3D CNN on Kinetics for Action Recognition

Multi-modal Transformer for Video Retrieval, ECCV 2020

Modeling Multi-Channel Videos with Expert Features: MMT

+ Audio: **6.9↓** + Speech: **0.4↓**

7 Expert Features

- OCR
- Speech
- Face
- Audio
- Scene
- Appearance
- Motion

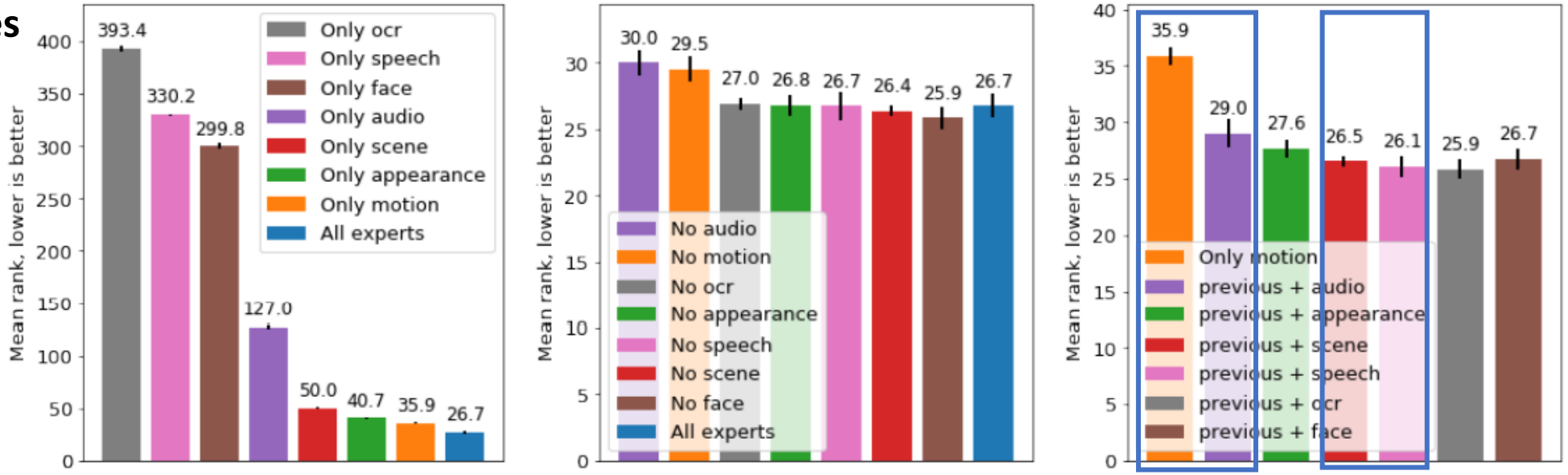


Fig. 4: MSRVTT performance (mean rank; lower is better) after training from scratch, when using only one expert (left), when using all experts but one (middle), when gradually adding experts by greedy search (right).

Text queries in MSRVTT are not specifically designed to describe multi-channel information in videos. Speech and audio expert features are helpful for retrieval performance.

Multi-modal Transformer for Video Retrieval, ECCV 2020

Modeling Multi-Channel Videos with Expert Features: MMT

+ Face: **0.8↑**

7 Expert Features

- OCR
- Speech
- Face
- Audio
- Scene
- Appearance
- Motion

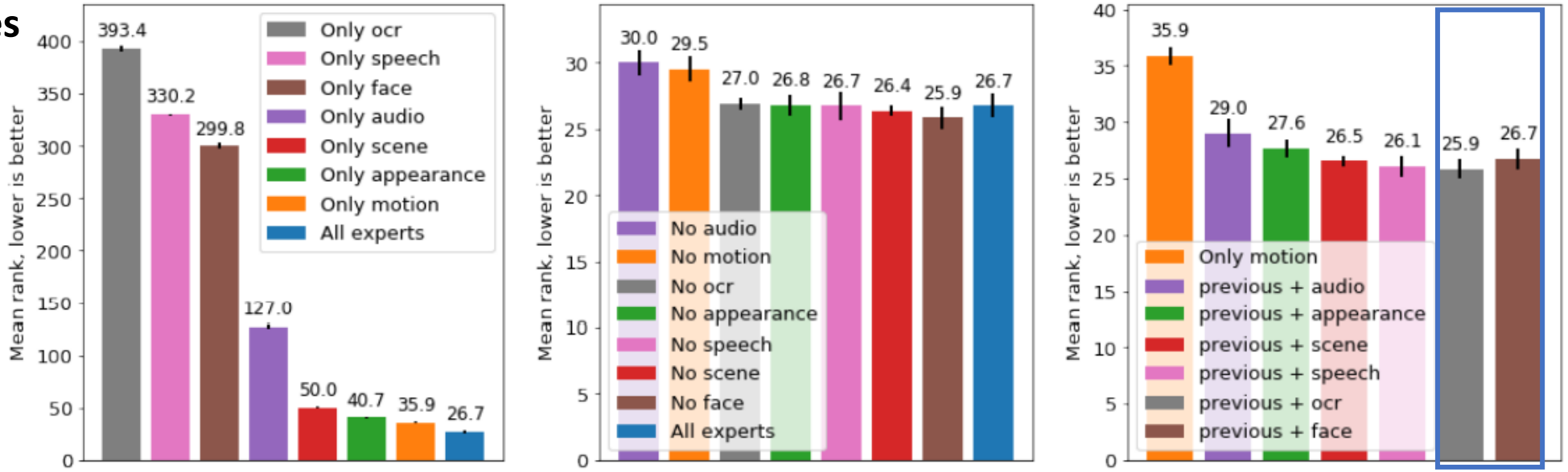
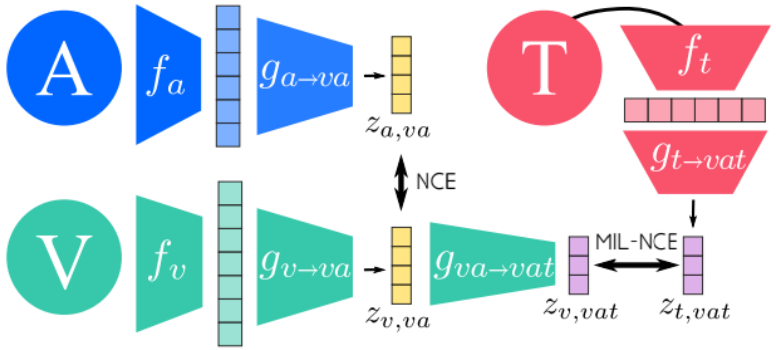


Fig. 4: MSRVTT performance (mean rank; lower is better) after training from scratch, when using only one expert (left), when using all experts but one (middle), when gradually adding experts by greedy search (right).

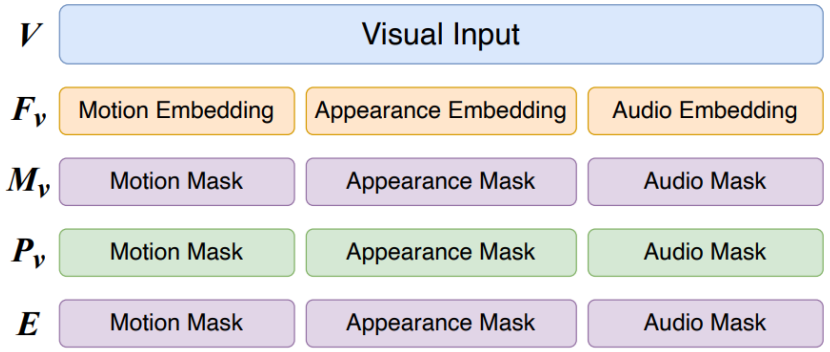
Not all visual features are helpful for downstream retrieval performance.

Multi-modal Transformer for Video Retrieval, ECCV 2020

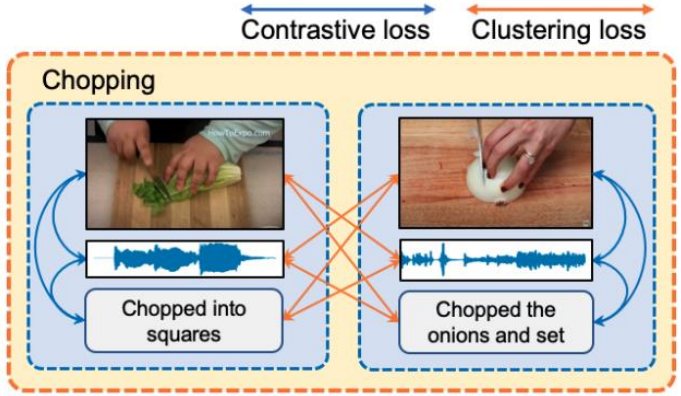
Modeling Multi-Channel Videos with Expert Features



Self-Supervised MultiModal Versatile Networks, NeurIPS 2020



HiT: Hierarchical Transformer With Momentum Contrast for Video-Text Retrieval, ICCV 2021

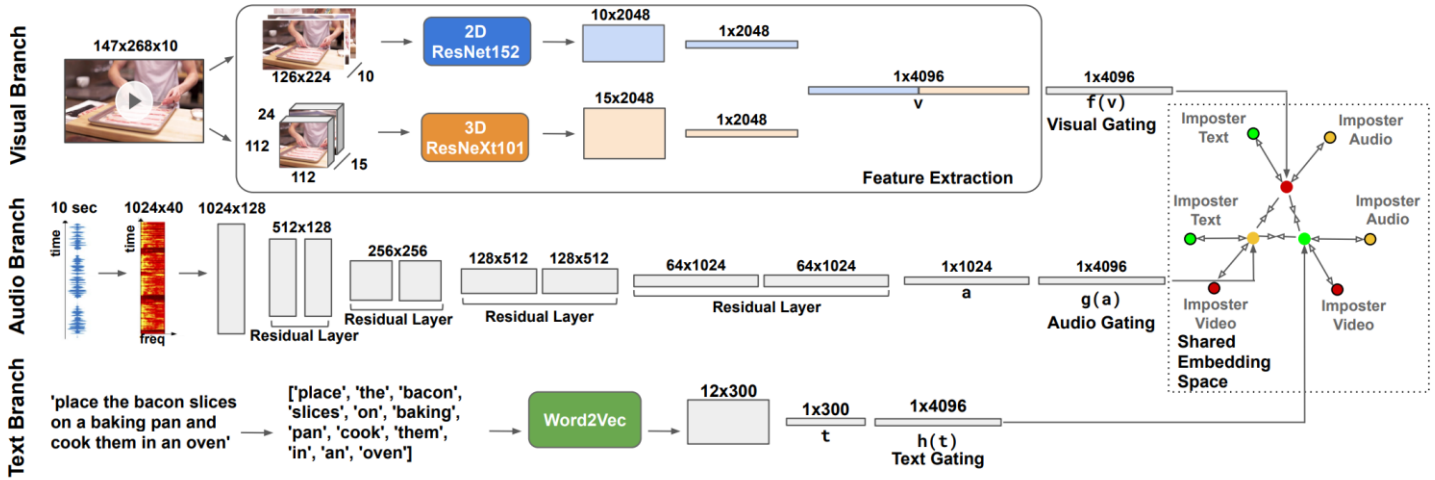


Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos, ICCV 2021

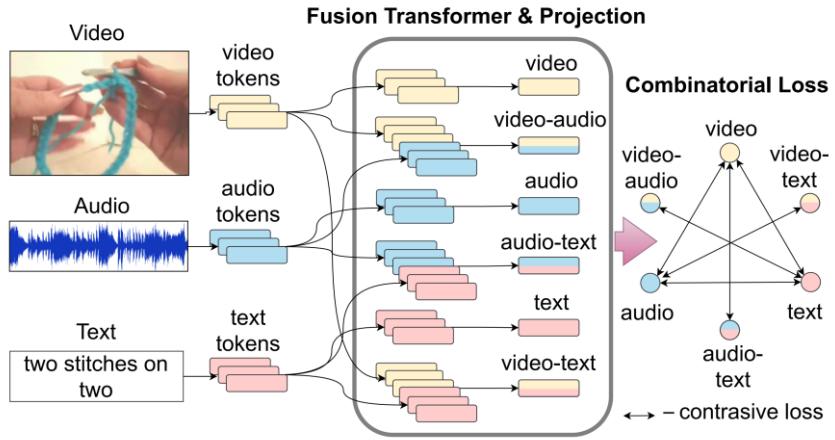
- Visual Expert Features
  - Motion: Pre-trained 3D CNN on Kinetics for Action Recognition
  - Appearance: Pre-trained 2D CNN on ImageNet for Image Classification
- Audio Expert Features
  - Pre-trained audio recognition models (e.g., on YT8M or AudioSet)
- Downstream Applications
  - Video-Text Retrieval (and Video Understanding)

Pre-extracted features from expert models, which are often well-supervised.

Modeling Multi-Channel Videos with Expert Features



AVLnet: Learning Audio-Visual Language Representations from Instructional Videos, Interspeech 2021



Everything at Once – Multi-modal Fusion Transformer for Video Retrieval, CVPR 2022

- Visual Expert Features
  - Motion: Pre-trained 3D CNN on Kinetics for Action Recognition
  - Appearance: Pre-trained 2D CNN on ImageNet for Image Classification
- ~~Audio Expert Features~~
  - ~~Pre-trained audio recognition models (e.g., on YT8M or AudioSet)~~
- Downstream Applications
  - Video-Text Retrieval and Audio-Video Retrieval

Random initialized audio encoder, learned end-to-end through self-supervised multi-modal pre-training

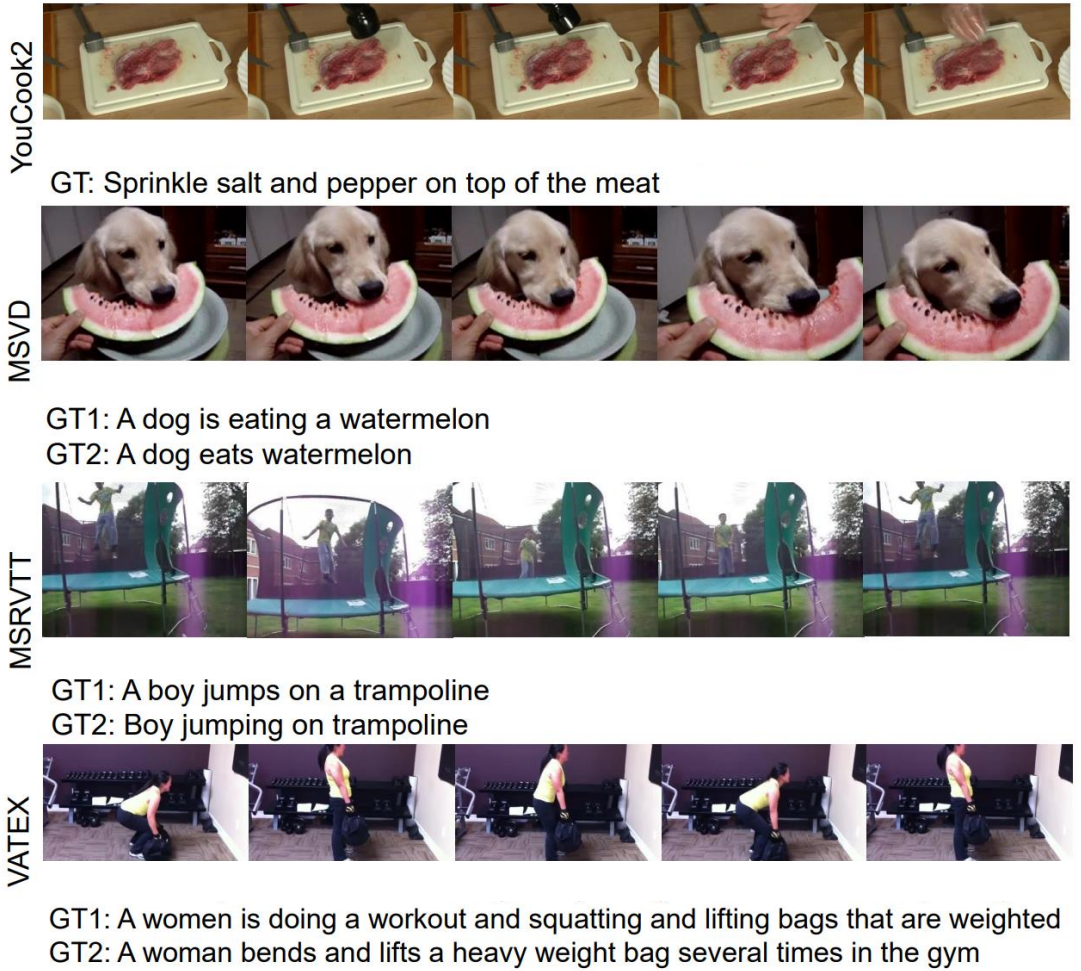
(Single-channel) Video+Language Tasks

Data Collection

Video

Task: Write a text description about the video

Sentence



Examples from SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning, CVPR 2022

Multi-channel Video+Language Tasks

00:00.755 --> 00:02.655
(Chandler:) Go to your room!
00:06.961 --> 00:08.622
(Janice:) I gotta go, I gotta go.

00:08.829 --> 00:10.057
(Janice:) Not without a kiss.
00:10.264 --> 00:12.391
(Chandler:) Maybe I won't kiss you so you'll stay.

00:12.600 --> 00:14.761
(Joey:) Kiss her. Kiss her!
00:16.771 --> 00:19.137
(Janice:) I'll see you later, sweetie. Bye, Joey.

...

00:39.327 --> 00:40.760
(Chandler:) She makes me happy.
00:41.596 --> 00:44.087
(Joey:) Okay. All right.

...

00:00 00:06 00:10 00:17 00:39 00:45 01:04

What is Janice holding on to after Chandler sends Joey to his room?

A Chandler's tie
B Chandler's hands
C Her Breakfast
D Her coat
E Chandler's coffee cup.

Why does Joey want Chandler to kiss Janice when they are in the kitchen?

A Because Joey is glad that Chandler is happy
B Because Joey likes to watch people kiss
C Because then she will leave
D Because Joey thinks Janice is hot
E Because then Chandler will move away from the toast.

What is on the couch behind Joey when he is at the counter?

A A chick
B A soccer ball
C A duck
D A pillow
E Janice's coat

TVQA: Localized, Compositional Video Question Answering, EMNLP 2018

Multi-channel Video+Language Tasks

TVR: Video-Subtitle Moment Retrieval

TVC: Multimodal Video Captioning

Castle : I'm so sorry for everything.
Mia: Come on, I did some pretty extraordinary things yesterday.

Captions

- Castle passes the flowers to Mia and Mia takes them. (*video-only*)
- Castle apologizes to the woman while handing her flowers. (*video-text*)

Ted: Just not on a boat.
Captain: Fair enough.

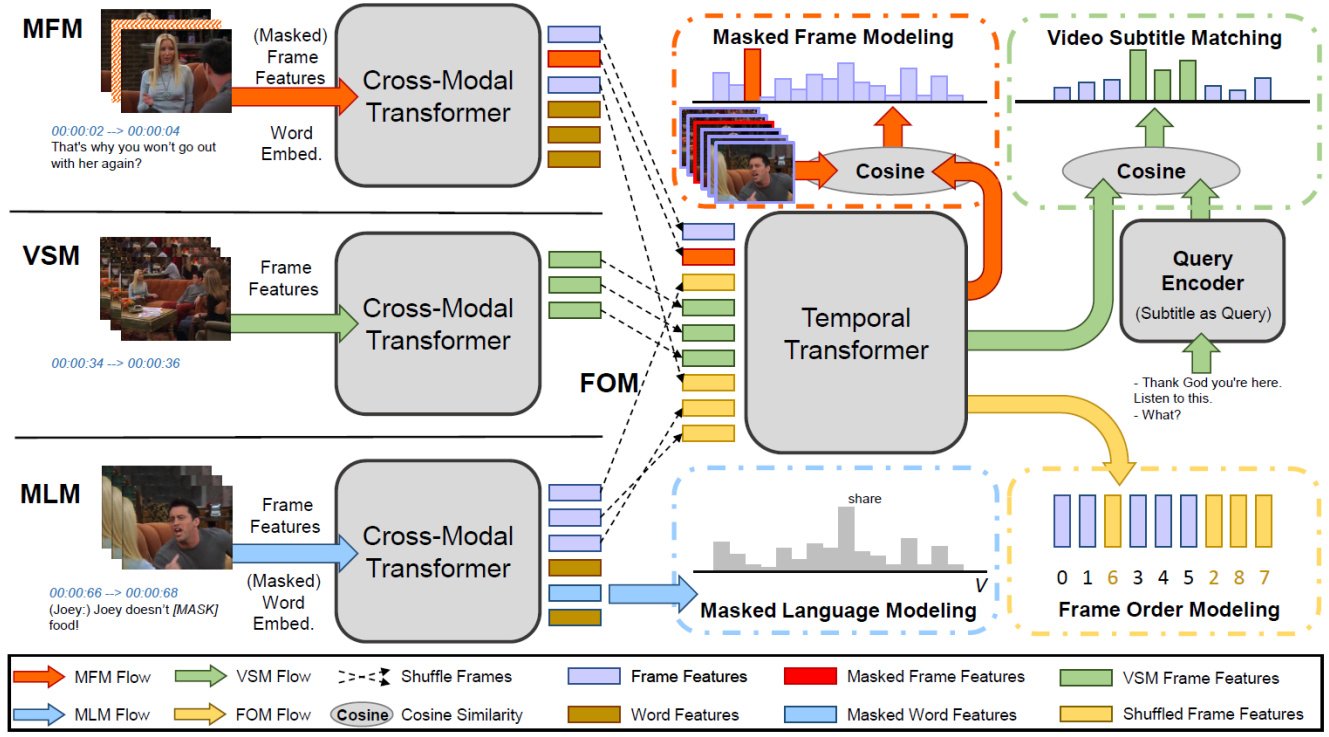
Captions

- The Captain says its ok if Ted will not be on the ship. (*text-only*)
- The Captain agrees and points at Ted with a glass in his hand. (*video-text*)

TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval, ECCV 2020

VLP for Multi-Channel Video+Language: HERO

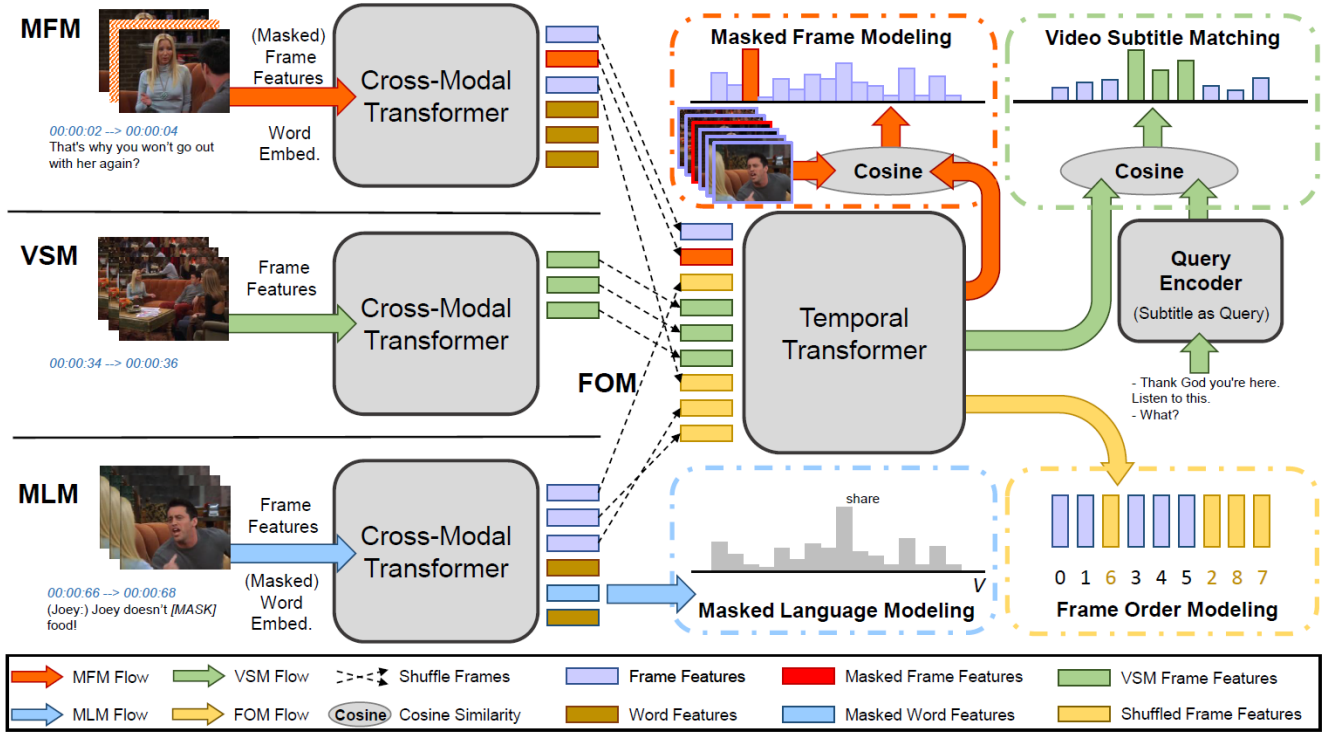
- Architecture
  - **Cross-Modal Transformer** for local temporal alignments between frames and subtitle sentences
  - **Temporal Transformer** for global temporal context



HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, EMNLP 2020

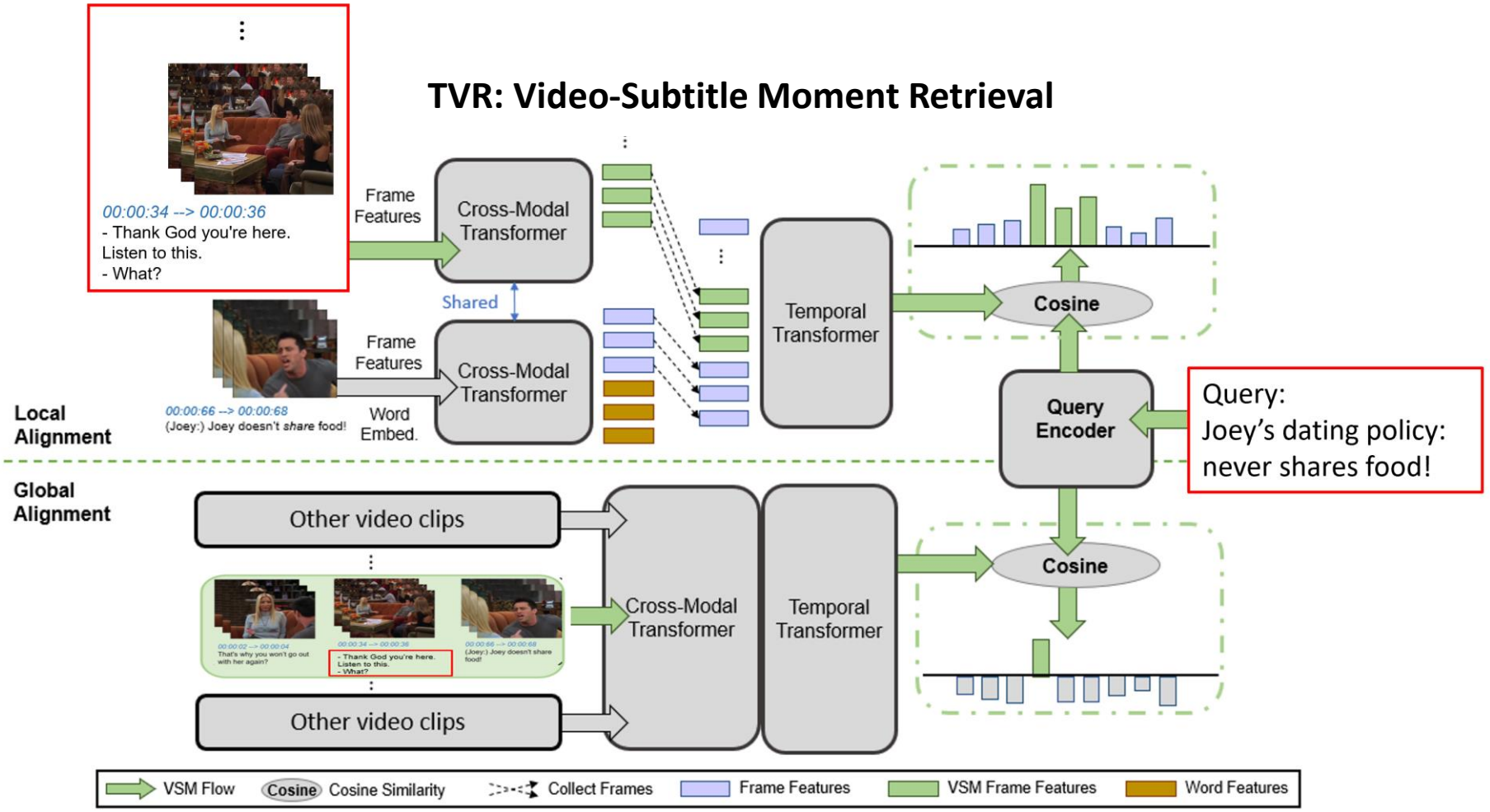
VLP for Multi-Channel Video+Language: HERO

- Architecture
  - **Cross-Modal Transformer** for local temporal alignments between frames and subtitle sentences
  - **Temporal Transformer** for global temporal context
- Pre-training Tasks
  - **Video Subtitle Matching** to learn both global and local alignment between the sampled query and video clips
  - **Frame Order Modeling** to reconstruct the orders of shuffled frames



HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, EMNLP 2020

VLP for Multi-Channel Video+Language: HERO



HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, EMNLP 2020

VLP for Multi-Channel Video+Language:HERO

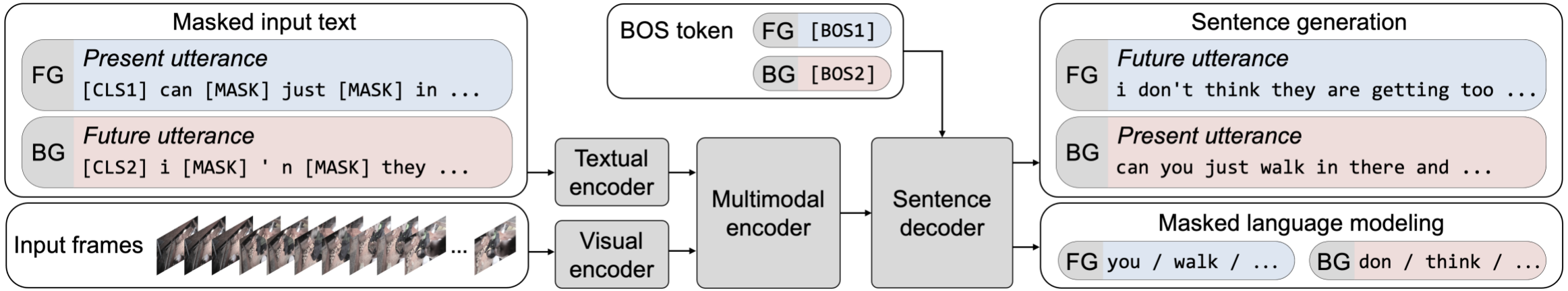
- Generalization to Single-channel Videos

Method \Task	DiDeMo			DiDeMo w/ ASR			MSR-VTT			MSR-VTT w/ ASR		
	R@1	R@10	R@100	R@1	R@10	R@100	R@1	R@5	R@10	R@1	R@5	R@10
SOTA Baseline	1.59	6.71	25.44	-	-	-	14.90	40.20	52.80	-	-	-
HERO	**2.14**	**11.43**	**36.09**	**3.01**	**14.87**	**47.26**	**16.80**	**43.40**	**57.70**	**20.50**	**47.60**	**60.90**

When augmenting single-channel videos with ASR inputs, the performance is improved

HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, EMNLP 2020

VLP for Multi-Channel Video+Language: MV-GPT



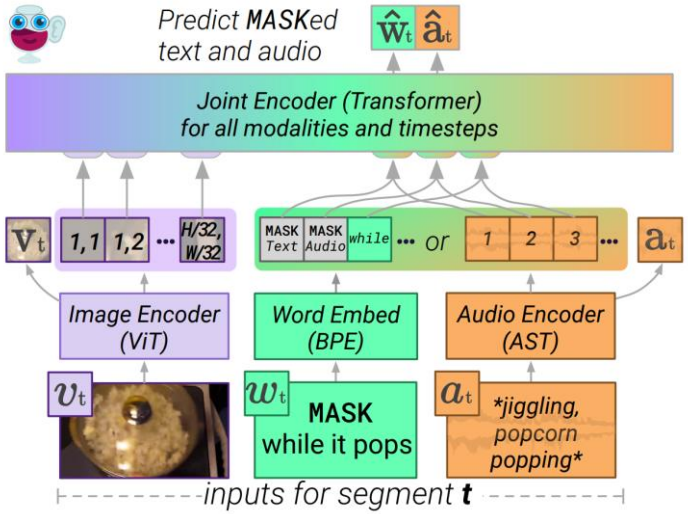
- Multimodal Video Generative Pre-training with Bi-directional Utterance Generation
  - **Forward Generation**: input frames + present utterance -> future utterance
  - **Backward Generation**: input frames + future utterance -> present utterance
- A pure Transformer-based architecture, end2end trained

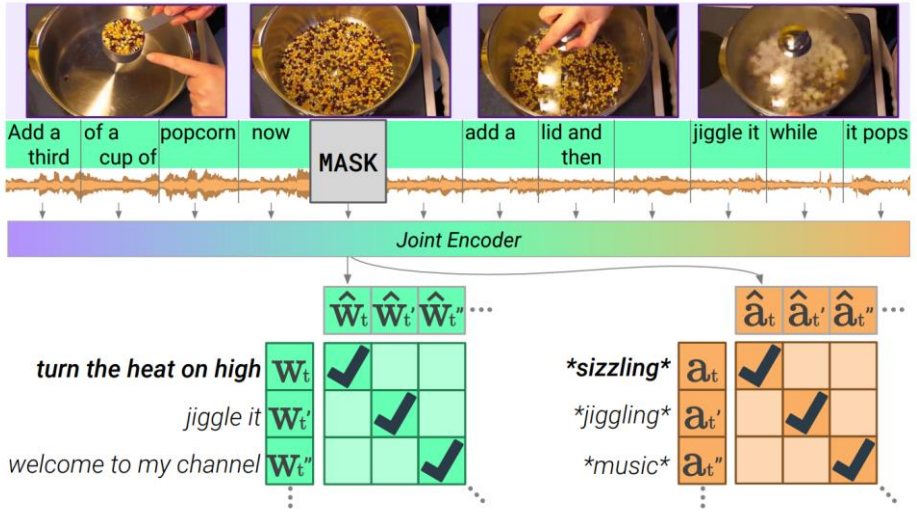
End-to-end Generative Pretraining for Multimodal Video Captioning, CVPR 2022

VLP for Multi-Channel Video+Language: MERLOT Reserve

- A Pure Transformer-based Architecture
  - Image Encoder: ViT
  - Audio Encoder: Audio Spectrum Transformer
  - Text/Joint Encoder: Transformer

- Pre-training on **YT-Temporal 1B** data
  - **Contrastive Masked Span Matching** on text and audio modalities
  - Contrastive matching between transcripts and frames
  - Trained in an end2end manner from scratch





MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound, CVPR 2022

VLP for Multi-Channel Video+Language : MERLOT Reserve

	Model	VCR test (acc; %) Q→A	QA→R	Q→AR
Caption/ObjDet-based	ERNIE-ViL-Large [124]	79.2	83.5	66.3
	Villa-Large [39]	78.9	83.8	65.7
	UNITER-Large [21]	77.3	80.8	62.8
	Villa-Base [39]	76.4	79.1	60.6
	VilBERT [81]	73.3	74.6	54.8
	B2T2 [4]	72.6	75.7	55.0
	VisualBERT [77]	71.6	73.2	52.4
Video-based	MERLOT [128]	80.6	80.4	65.1
	RESERVE-B	79.3	78.7	62.6
	RESERVE-L	**84.0**	**84.9**	**72.0**

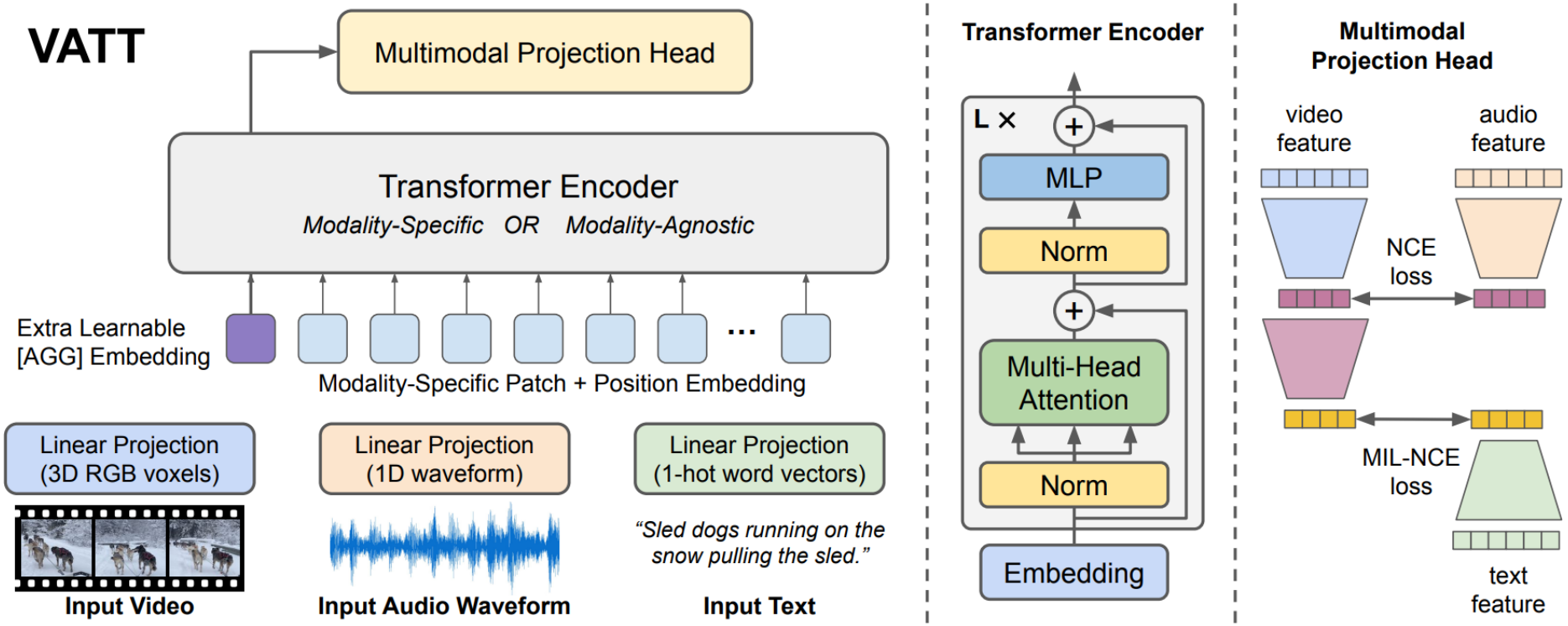
Video-Text Pre-training can help Visual Commonsense Reasoning, an image-text task

	Configuration *for one epoch of pretraining*	VCR Q→A	val (%)
V+T	Mask LM [29, 106, 128]	67.2	
	VirTex-style [27]	67.8	
	Contrastive Span	**69.7**	
V+T+A	Audio as target	70.4	
	Audio as input and target	**70.7**	
	Audio as input and target, w/o strict localization	70.6	
	RESERVE-B	**71.9**	

Audio pre-training helps, even for the audio-less VCR

MERLOT: Multimodal Neural Script Knowledge Models, NeurIPS 2021
MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound, CVPR 2022

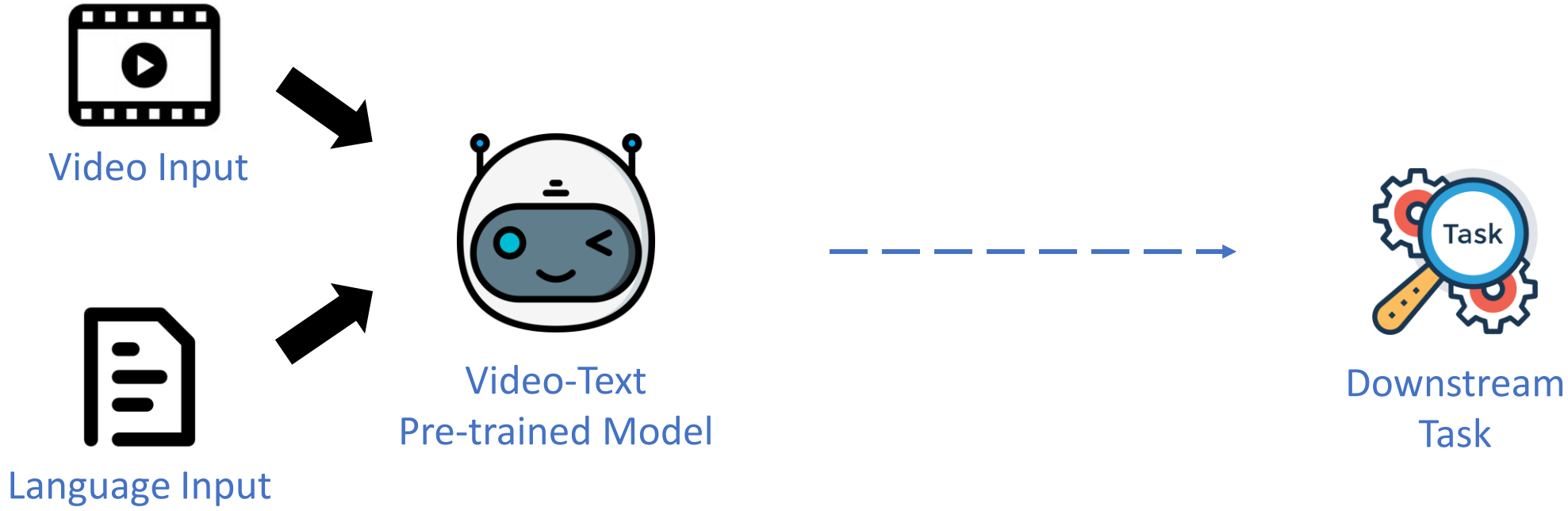
Unified Architecture for Multi-Channel Video Encoding: VATT



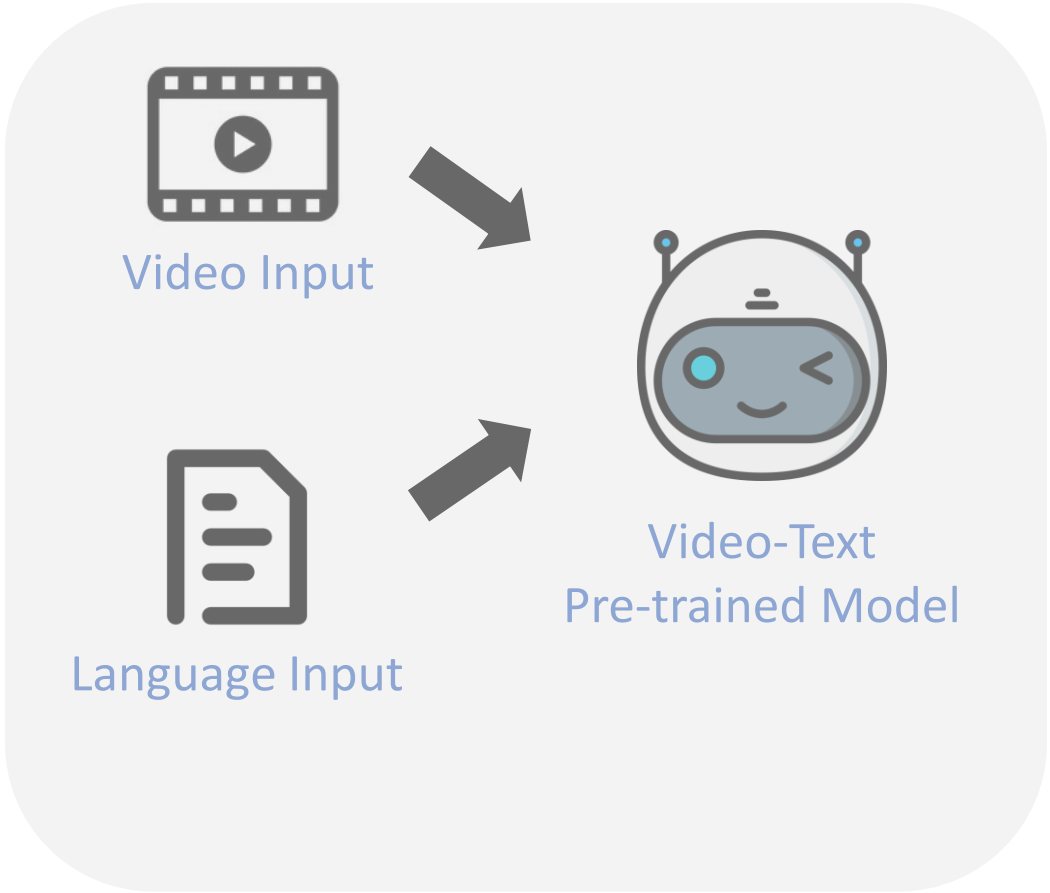
A plain Transformer architecture for all modalities.

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text, NeurIPS 2021

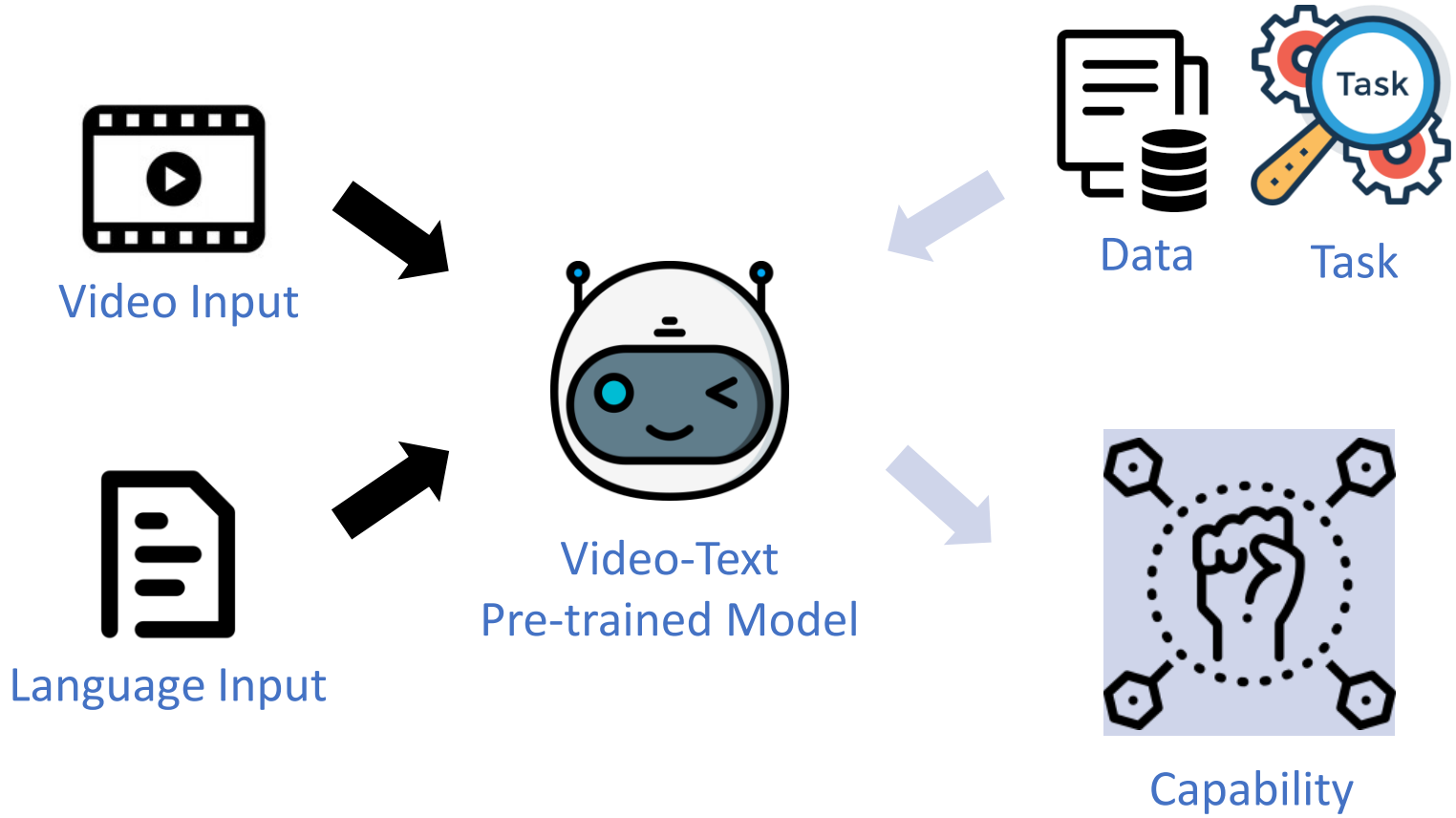
Benchmarking VidL Models



Benchmarking VidL Models



Benchmarking VidL Models



Benchmarking VidL Models

Capability 1: A general VidL system should do well on diverse tasks/domains/datasets.

Method	Retrieval Tasks	QA Tasks	Captioning Tasks
HowTo100M [Miech et al.]	MSRVTT, YouCook2	-	-
HERO [Li et al.]	MSRVTT, DiDeMo, TVR, How2R	TVQA, How2QA	TVC
ActBERT [Zhu and Yang]	MSRVTT, YouCook2	MSRVTT, LMSDC	YouCook2
ClipBERT [Lei et al.]	MSRVTT, ActivityNet, DiDeMo	TGIF, MSRVTT	-
VQA-T [Yang et al.]	-	MSRVTT, MSVD, How2QA, ….	-
Frozen in Time [Bain et al.]	MSRVTT, MSVD, DiDeMo, LSMDC	-	-
VIOLET [Fu et al.]	MSRVTT, MSVD, YouCook2, LSMDC	TGIF, MSRVTT, MSVD, LSMDC	-
MV-GPT [Seo et al.]	MSRVTT	MSRVTT, ActivityNet	YouCook2, MSRVTT, …
MERLOT [Zeller et al.]	-	TGIF, MSRVTT, LSMDC, TVQA,…	-
MERLOT RESERVE [Zeller et al.]	-	TVQA, MSRVTT, MSVD, …	-

Benchmarking VidL Models

Capability 2: A smart VidL system should be able to leverage information from different modalities in video.

Video Frames

Single-Channel Video

V.S.

Video Frames

ASR/Subtitles

Audio

Multi-Channel Video

Method	MSRVTT-QA (Single-channel)	TVQA (Multi-channel)
HERO [Li et al.]	-	73.6
DECEMBERT [Tan et al.]	37.4	-
ClipBERT [Lei et al.]	37.4	-
SiaSamRea [Lei et al.]	41.6	-
VQA-T [Yang et al.]	41.5	-
MERLOT [Zeller et al.]	43.1	78.7
VIOLET [Fu et al.]	43.7	-
MV-GPT [Seo et al.]	41.7	-
MERLOT RESERVE [Zeller et al.]	-	86.5

Benchmarking VidL Models

NLP Benchmarks

XTREME

(X) Cross-Lingual **T**ransfer **E**valuation of **M**ultilingual **E**ncoders

Publicly accessible large-scale multi-task benchmarks can facilitate advances in modeling.

VALUE Benchmark https://value-benchmark.github.io

- A comprehensive benchmark for **V**ideo-**A**nd-**L**anguage **U**nderstanding **E**valuation

Multi-channel Video

With both **Video Frames** and **Subtitle/ASR**

Diverse Video Domain

Diverse video content from **YouTube**, **TV Episodes** and **Movie Clips**

Various Datasets over Representative Tasks

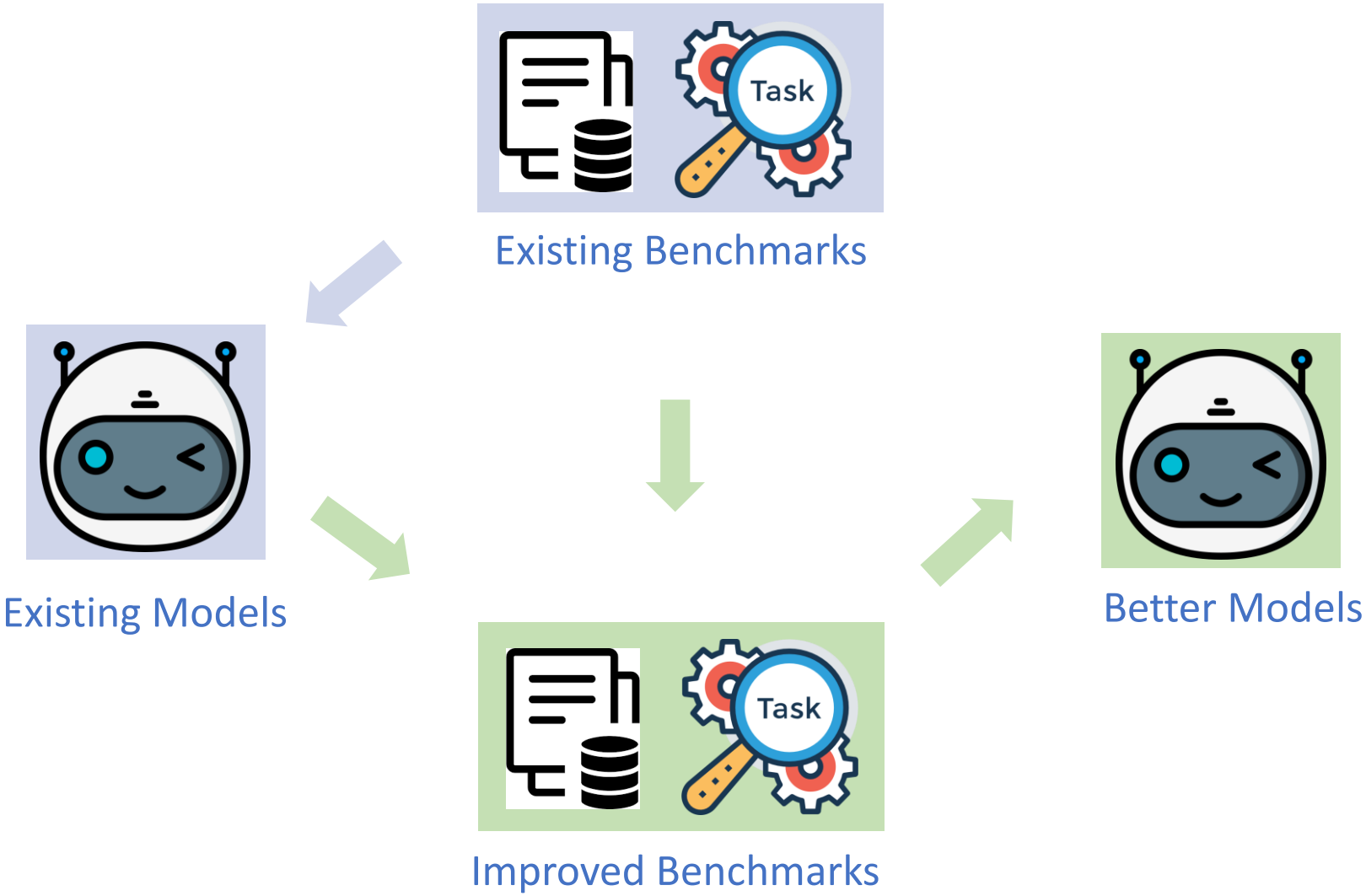
11 datasets over **3** tasks: **Retrieval**, **Question Answering** and **Captioning**.

Leaderboard!

To track the advances in Video-and-Langauge research.

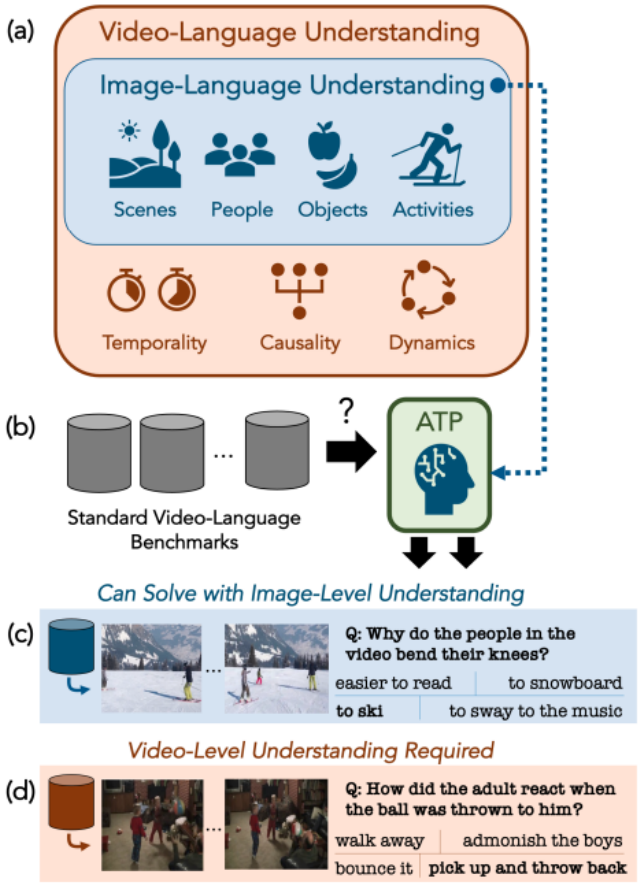
VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation, NeurIPS 2021 Datasets and Benchmarks Track

Analyzing VidL Tasks and Models



Analyzing VidL Tasks

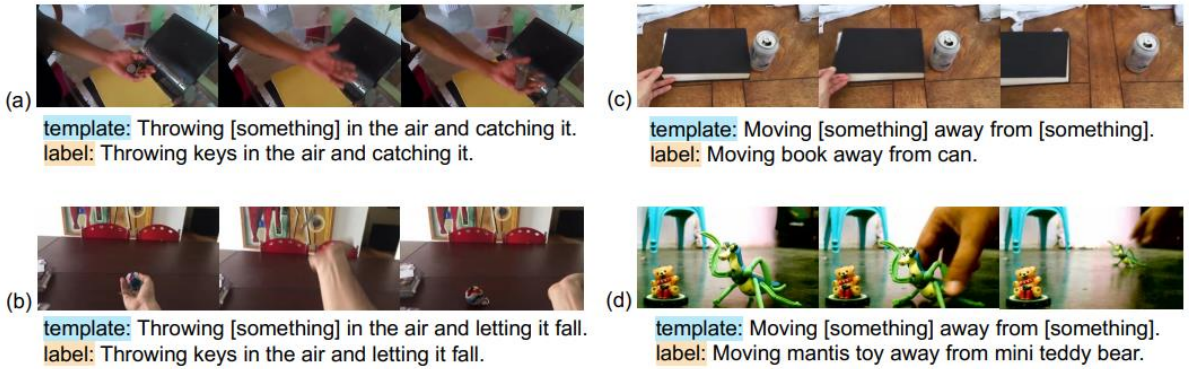
Do existing video-language tasks require temporal reasoning?



Revisiting the "Video" in Video-Language Understanding, CVPR 2022

Method	#PT	#Train Frame	MSRVTT R1	MSRVTT R5	MSRVTT R10	DiDeMo R1	DiDeMo R5	DiDeMo R10	ActivityNet Cap R1	ActivityNet Cap R5	ActivityNet Cap R10
HERO [37]	136M	310	20.5	47.6	60.9	-	-	-	-	-	-
ClipBERT [31]	0.2M	16/16/8	22.0	46.8	59.9	20.4	48.0	60.8	21.3	49.0	63.5
VideoCLIP [61]	136M	960	30.9	55.4	66.8	-	-	-	-	-	-
Frozen [4]	5M	4	31.0	59.5	70.5	31.0	59.8	72.4	-	-	-
AlignPrompt [34]	5M	8	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
All-in-one [58]	138M	9	34.4	65.4	75.8	32.7	61.4	73.5	22.4	53.7	67.7
CLIP4Clip [47]	400M	12/64/64	42.0	68.6	78.7	42.8	68.5	79.2	40.5	72.4	98.2
SINGULARITY	5M	1	36.8	65.9	75.5	47.4	75.2	84.0	43.0	70.6	81.3
SINGULARITY	17M	1	**41.5**	**68.7**	**77.0**	**53.9**	**79.4**	**86.9**	**47.1**	**75.5**	**85.5**

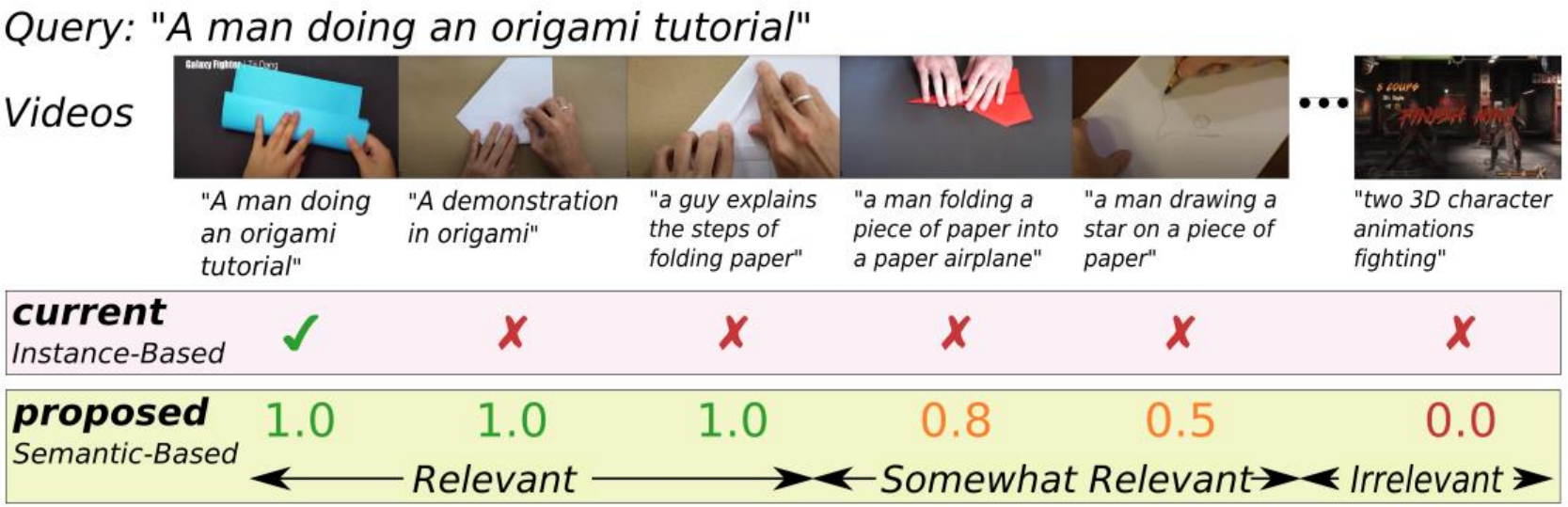
SSV2-Template/Label Retrieval



Revealing Single Frame Bias for Video-and-Language Learning, arxiv 2022

Analyzing VidL Tasks

Shortcomings in current video retrieval benchmarks and evaluation protocols

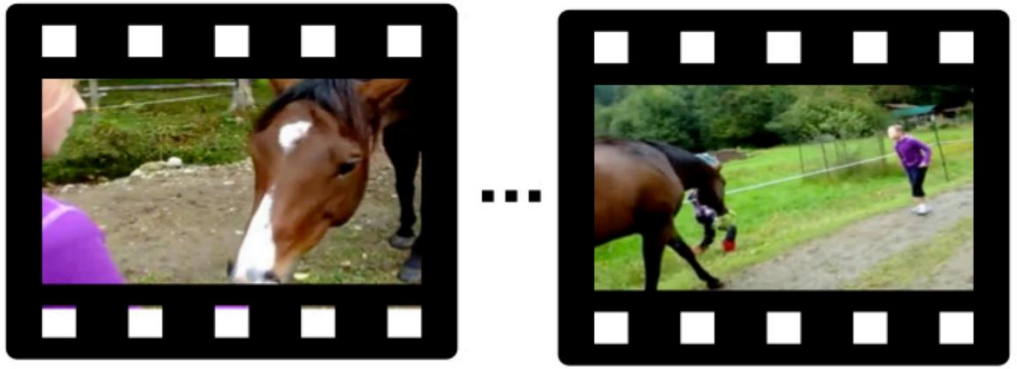


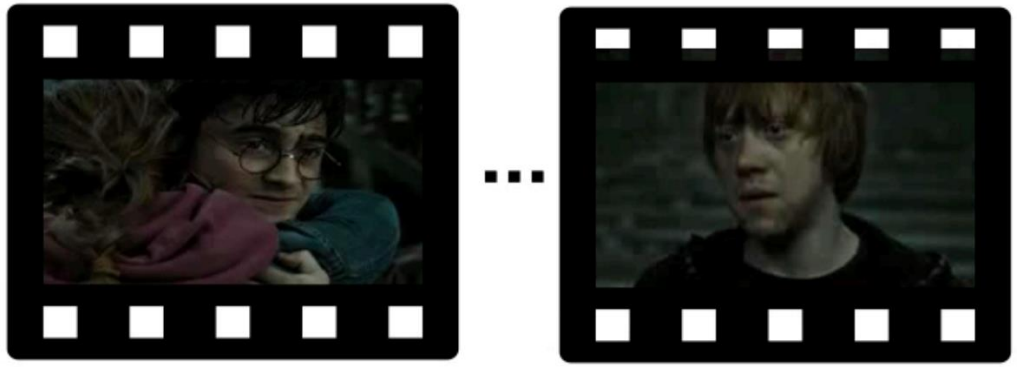
- Assumptions in existing video retrieval benchmarks
  - Only a single caption is relevant to a query video and vice versa
- New evaluation protocol and training objective
  - Semantic similarity based on proxy measures with Bag of Words, Part of Speech, Synset, METEOR.

On Semantic Similarity in Video Retrieval, CVPR 2021

Analyzing VidL Models

When do existing video-language models fail?





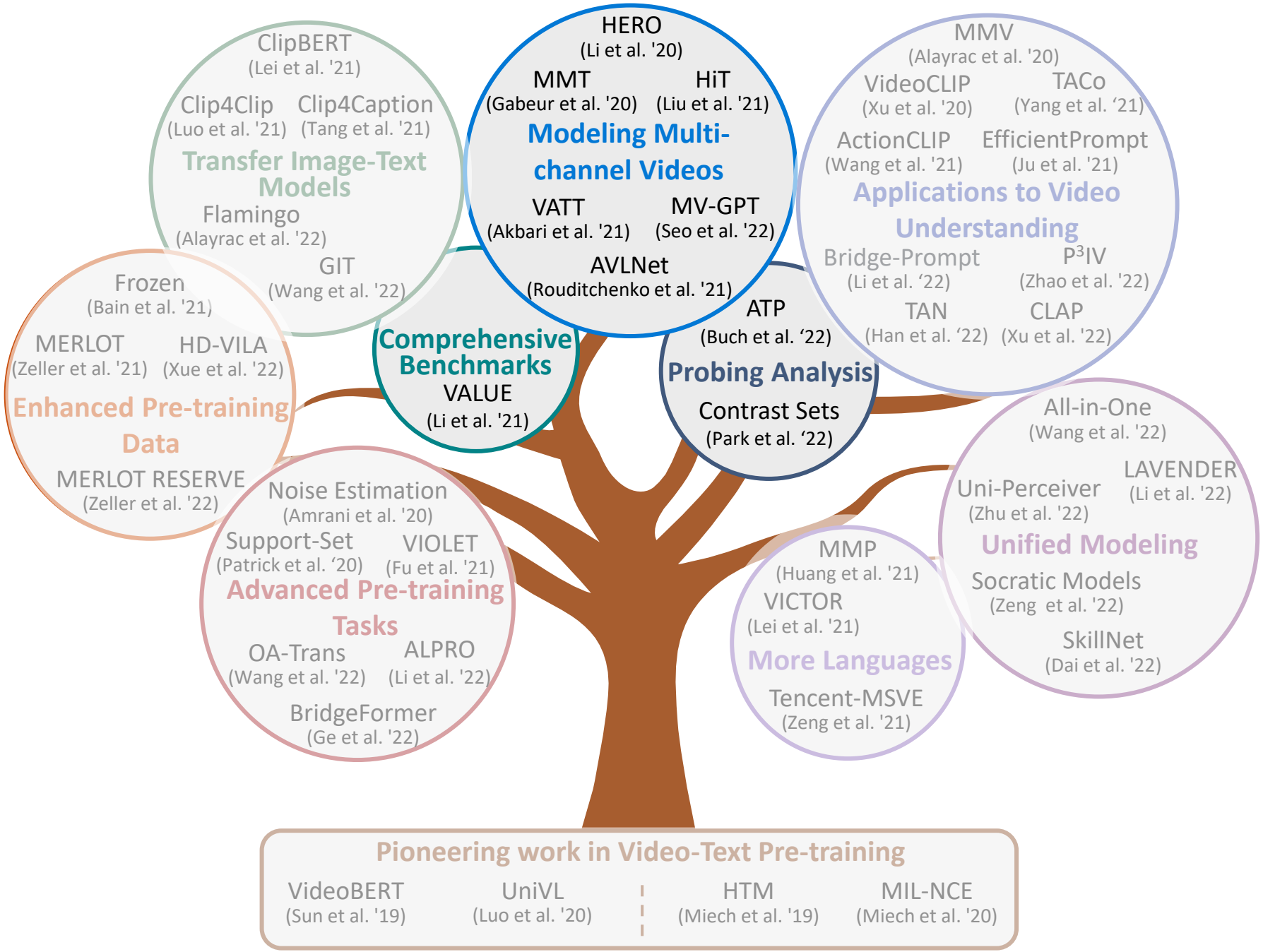
Verb Manipulation

A: A girl feeding a brown horse. **GT**
B: A girl rides a brown horse. **Predicted**
C: Football team playing football on a field.
D: The man is drinking beer.
E: Two men playing a video game.

Entity Manipulation

A: Over her shoulders, HARRY glances at RON, who lowers his gaze for a moment. **GT**
B: Over her shoulders RON glances at HARRY, who lowers his gaze for a moment. **Predicted**
C: He tries to shake him off.
D: HARRY studies it.
E: RON whispers to HARRY.

Exposing the Limits of Video-Text Models through Contrast Sets, NAACL Short 2022


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

Looking forward

- How to effectively leverage advances in unimodal models
  - E2E pre-training > pre-extracted expert features
  - E2E pre-training (from scratch) is usually time-consuming, extremely data-hungry, very computationally expensive

- Benchmarks on truly multi-channel video understanding
  - Existing VidL datasets do not test on audio understanding in video