JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu,
Ce Liu, Lijuan Wang

Paper Tag: THU-PM-240

LAVENDER: unify all as open-vocabulary generation via MLM

-> Removes task-specific heads, all task can share the same MLM head

-> Can easily adapted to multi-task finetuning

-> Enable zero-shot capability on QA tasks, even without leveraging the super power from LLMs

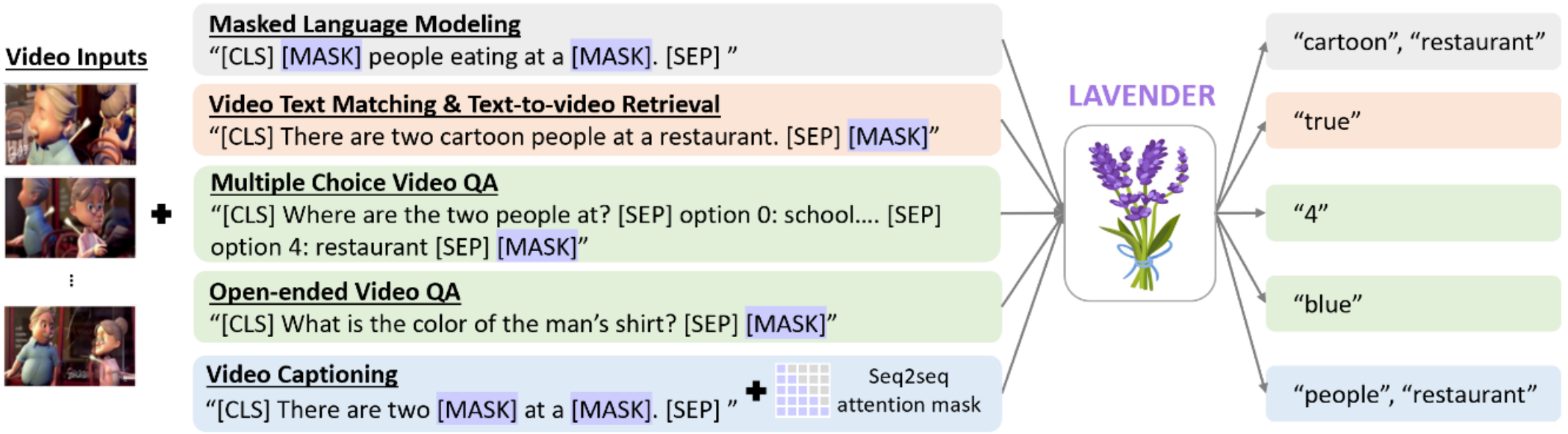


Figure 1. Overview of LAVENDER (**LA**nguage-**V**id**E**o u**NDER**standing) model. LAVENDER unifies both pre-training and downstream finetuning as Masked Language Modeling.

Common practices in Video-language Modeling

-> Add a task-specific head for each task or even each dataset

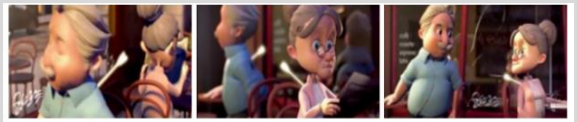
-> No ZS capability for QA tasks

Video QA

What is a cute model doing? → *Walking*

[Open-ended] **Classification** over a pre-defined answer dictionaries
[Multiple-choice] **Classification** over the answer choices

Video Captioning

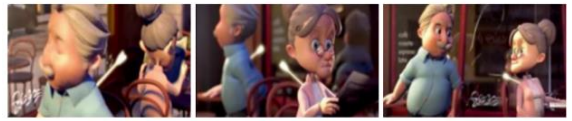


→ *"cartoon people eating at restaurant"*

Open-vocabulary **generation**

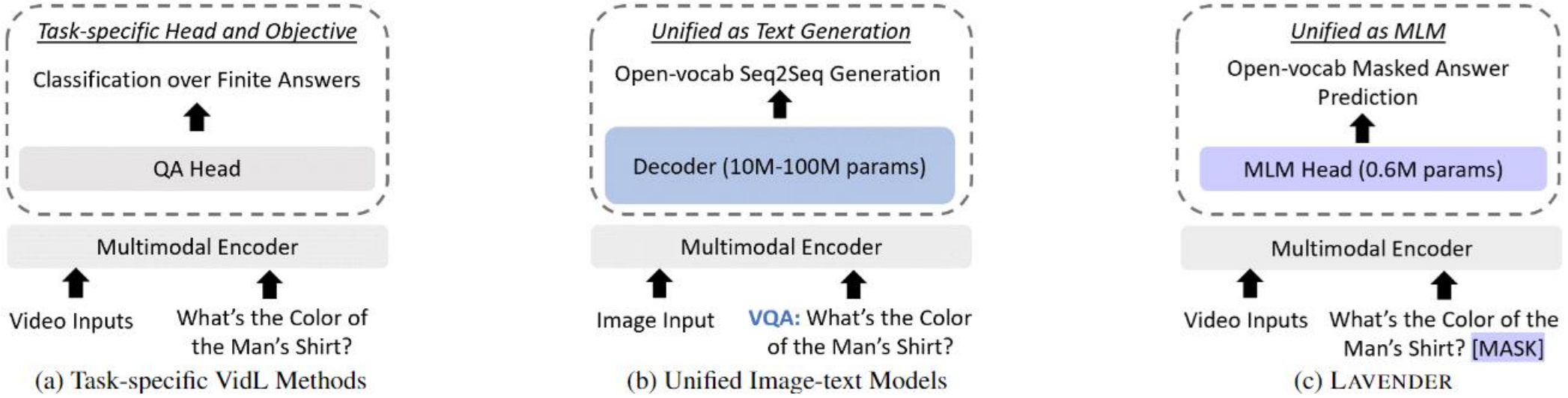
Text-to-video Retrieval

"cartoon people eating at restaurant" →



Classification / **Ranking** over positive pairs and negative pairs

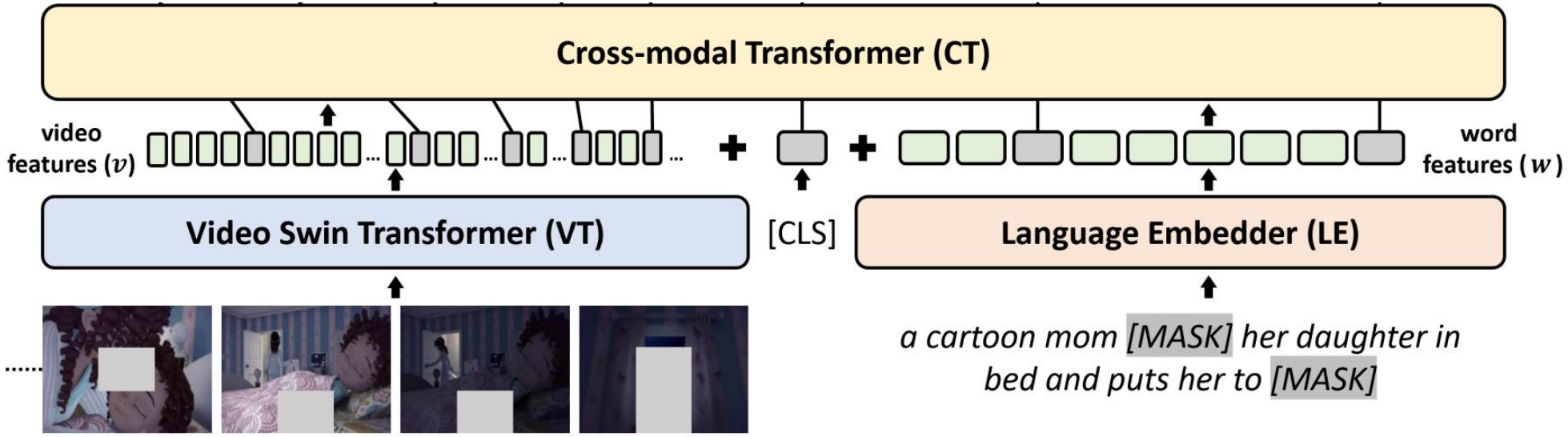
Comparison to existing methods



- Unlike task-specific designs in existing VidL methods, LAVENDER unifies all tasks as MLM
- We adopt an encoder-only architecture, with a lightweight MLM head, instead of the heavy decoder in unified image-text models

LAVENDER

- Model Architecture
  - Text Encoder: word embedding layer
  - Video Encoder: Video Swin Transformer
  - Fusion Encoder: 12 Transformer layers for cross-modal modeling

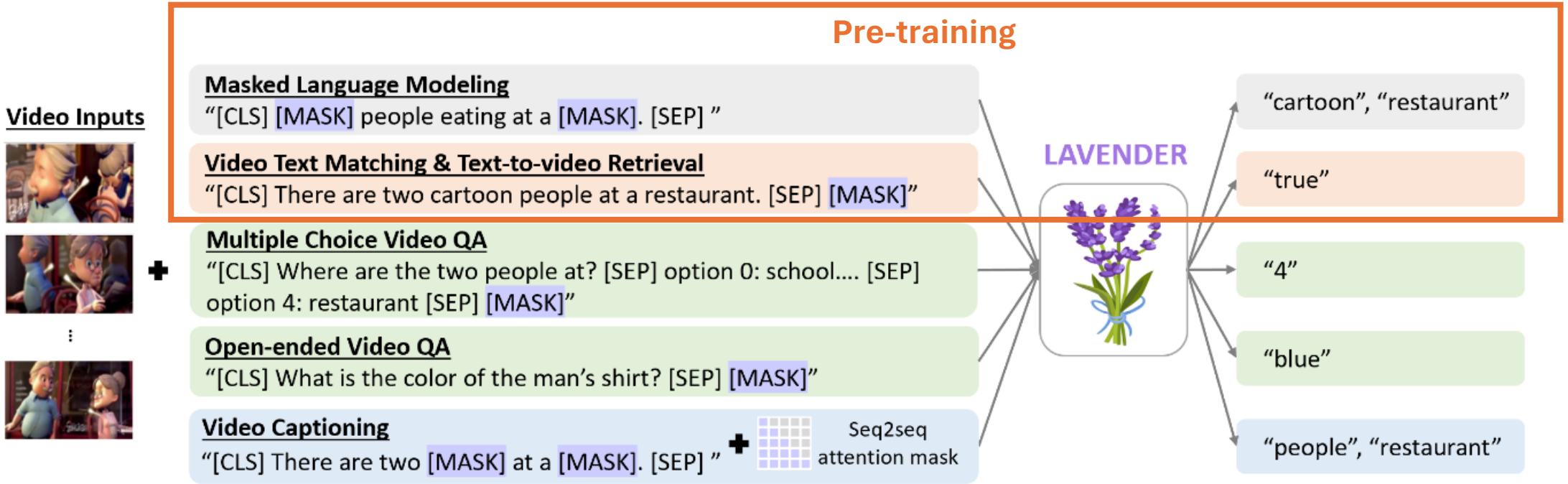


VIOLET

VIOLET : End-to-End Video-Language Transformers with Masked Visual-token Modeling, arXiv 2021

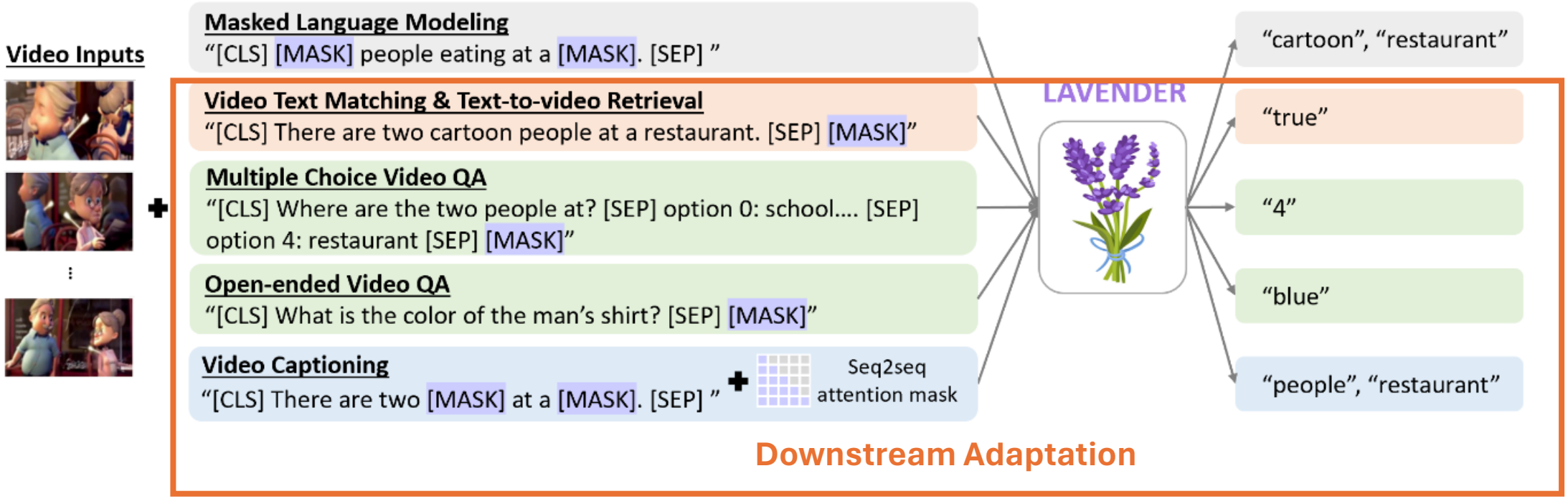
LAVENDER

- Model Architecture
  - Text Encoder: word embedding layer
  - Video Encoder: Video Swin Transformer
  - Fusion Encoder: 12 Transformer layers for cross-modal modeling



LAVENDER

- Model Architecture
  - Text Encoder: word embedding layer
  - Video Encoder: Video Swin Transformer
  - Fusion Encoder: 12 Transformer layers for cross-modal modeling



Comparison to task-specific baseline

VidL Pre-training	Task-specific designs	Finetune setting	#Params	#	Meta Ave.	TGIF Action	MSVD QA	DiDeMo Ret.	MSRVTT Cap.
-	-	ST	4(P+H)	1	45.5	93.5	40.8	0.0[4]	47.7
		MT	P+H	2	58.5	95.9	47.4	41.2	50.0
	Head	ST	4(P+H)	3	40.1	31.9	44.2	36.7	47.4
		MT	P+4H	4	55.6	94.1	44.6	35.4	48.3
VTM+MLM	Head	ST	4(P+H)	5	64.0	94.5	46.7	59.0	55.7
		MT	P+4H	6	62.4	95.5	47.7	53.0	53.3
VTM (as MLM)+MLM	-	ST	4(P+H)	7	**68.9**	95.8	**54.4**	**68.2**	57.3
	-	MT	P+H	8	68.3	**96.5**	53.5	65.8	**57.4**
	Task Prompt			9	67.9	96.2	53.4	65.6	56.4
	Task Token			10	67.9	**96.5**	53.6	64.9	56.7

- Task-specific baseline with different head designs for different tasks vs. LANVENDER with the same MLM head for all tasks

Comparison to task-specific baseline (w/ video-language pre-training)

VidL Pre-training	Task-specific designs	Finetune setting	#Params	#	Meta Ave.	TGIF Action	MSVD QA	DiDeMo Ret.	MSRVTT Cap.
-	-	ST	4(P+H)	1	45.5	93.5	40.8	0.0[4]	47.7
		MT	P+H	2	58.5	95.9	47.4	41.2	50.0
	Head	ST	4(P+H)	3	40.1	31.9	44.2	36.7	47.4
		MT	P+4H	4	55.6	94.1	44.6	35.4	48.3
VTM+MLM	Head	ST	4(P+H)	5	64.0	94.5	46.7	59.0	55.7
		MT	P+4H	6	62.4	95.5	47.7	53.0	53.3
VTM (as MLM)+MLM	-	ST	4(P+H)	7	**68.9**	95.8	**54.4**	**68.2**	57.3
	-	MT	P+H	8	68.3	**96.5**	53.5	65.8	**57.4**
	Task Prompt			9	67.9	96.2	53.4	65.6	56.4
	Task Token			10	67.9	**96.5**	53.6	64.9	56.7

- Single-task Finetuning
  - LAVENDER (L5) significantly outperforms task-specific baseline (L7), with +4.9 on Meta-Ave.

Comparison to task-specific baseline (w/ video-language pre-training)

VidL Pre-training	Task-specific designs	Finetune setting	#Params	#	Meta Ave.	TGIF Action	MSVD QA	DiDeMo Ret.	MSRVTT Cap.
-	-	ST	4(P+H)	1	45.5	93.5	40.8	0.0[4]	47.7
		MT	P+H	2	58.5	95.9	47.4	41.2	50.0
	Head	ST	4(P+H)	3	40.1	31.9	44.2	36.7	47.4
		MT	P+4H	4	55.6	94.1	44.6	35.4	48.3
VTM+MLM	Head	ST	4(P+H)	5	64.0	94.5	46.7	59.0	55.7
		MT	P+4H	6	62.4	95.5	47.7	53.0	53.3
VTM (as MLM)+MLM	-	ST	4(P+H)	7	**68.9**	95.8	**54.4**	**68.2**	57.3
	-			8	68.3	**96.5**	53.5	65.8	**57.4**
	Task Prompt	MT	P+H	9	67.9	96.2	53.4	65.6	56.4
	Task Token			10	67.9	**96.5**	53.6	64.9	56.7

- Single-task Finetuning
  - LAVENDER (L5) significantly outperforms task-specific baseline (L7), with +4.9 on Meta-Ave.
- Multi-task Finetuning
  - LAVENDER (L6) consistently outperforms task-specific baseline (L8), with +5.9 on Meta-Ave.

Comparison to task-specific baseline (w/ video-language pre-training)

VidL Pre-training	Task-specific designs	Finetune setting	#Params	#	Meta Ave.	TGIF Action	MSVD QA	DiDeMo Ret.	MSRVTT Cap.
-	-	ST	4(P+H)	1	45.5	93.5	40.8	0.0[4]	47.7
		MT	P+H	2	58.5	95.9	47.4	41.2	50.0
	Head	ST	4(P+H)	3	40.1	31.9	44.2	36.7	47.4
		MT	P+4H	4	55.6	94.1	44.6	35.4	48.3
VTM+MLM	Head	ST	4(P+H)	5	64.0	94.5	46.7	59.0	55.7
		MT	P+4H	6	62.4	95.5	47.7	53.0	53.3
VTM (as MLM)+MLM	-	ST	4(P+H)	7	**68.9**	95.8	**54.4**	**68.2**	57.3
	-	MT	P+H	8	68.3	**96.5**	53.5	65.8	**57.4**
	Task Prompt			9	67.9	96.2	53.4	65.6	56.4
	Task Token			10	67.9	**96.5**	53.6	64.9	56.7

- Single-task Finetuning
  - LAVENDER (L5) significantly outperforms task-specific baseline (L7), with +4.9 on Meta-Ave.
- Multi-task Finetuning
  - LAVENDER (L6) consistently outperforms task-specific baseline (L8), with +5.9 on Meta-Ave.
  - LAVENDER can also support task-specific prompt (L9) / token (L10) for multi-task finetuning, by simply prepending the prompt or a learnable token to the text input, but does not bring performance improvements

Multi-task finetuning

Can we have a unified architecture that supports all downstream tasks simultaneously without introducing task-specific heads?

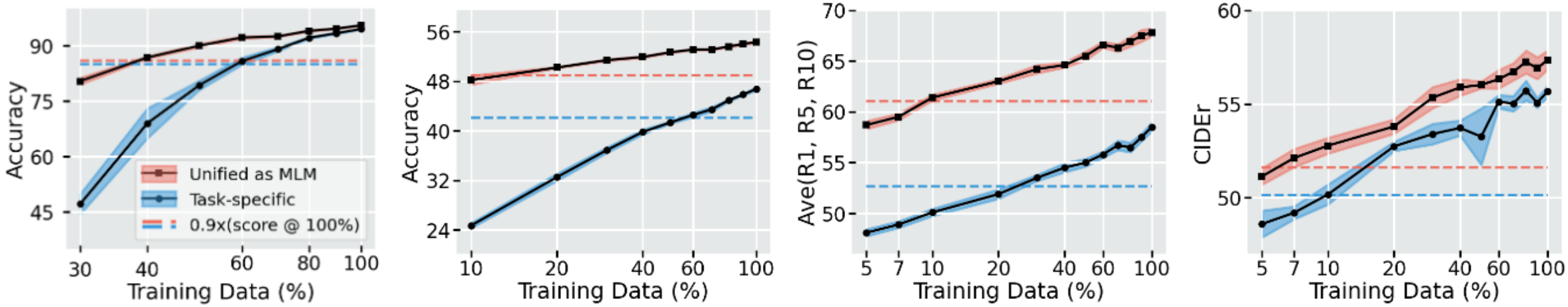
Multi-task Settings

- MT (all-in-one): a single set of parameters for all tasks
- MT (best): the best performing checkpoint for each task while training MT (all-in-one)
- MT -> ST: with multi-task finetuning as 2nd stage pre-training and then finetune on each task

Finetune Method	# Params	Meta Ave.	TGIF			MSRVTT				LSMDC			MSVD			DiDeMo
			Act.	Trans.	Frame	MC	QA	Ret	Cap	MC	FiB	Ret	QA	Ret	Cap	Ret
ST	14P	73.9	95.8	**99.1**	**72.2**	**96.6**	**44.2**	**58.9**	57.3	84.5	**56.9**	**39.8**	54.4	67.6	139.4	**68.2**
MT (all-in-one)	P	73.4	95.8	98.0	70.7	93.9	44.1	56.3	57.1	85.3	56.5	39.4	53.4	69.2	141.1	66.1
MT (best)	14P	73.8	95.8	98.3	71.6	94.3	**44.2**	56.4	57.2	**86.0**	56.7	39.4	**55.4**	69.3	141.6	66.5
MT → ST	14P	**74.2**	**96.6**	98.5	71.2	96.0	44.1	58.8	**58.0**	85.3	**56.9**	**39.8**	53.5	**69.7**	**142.9**	67.7
MT (all-in-one) TS	>P	69.2	93.8	97.2	65.4	92.2	41.7	52.7	54.2	83.0	49.5	34.7	49.2	65.6	133.7	56.5

- Best performing setting: MT -> ST
- All-in-one is very competitive, with only -0.5 performance drop from ST baseline on Meta Ave.
- Compared to task-specific baseline, we observe a consistent gain of +4.2 on Meta-Ave.

Few-shot Generalizability



(a) TGIF-Action (b) MSVD-QA (c) DiDeMo Retrieval (d) MSRVTT Captioning

Figure 3. **Few-shot Evaluation** under VidL Pre-training. Each experiment are repeated 5 times with different random seeds. The shaded areas highlight the standard error. Percentage of training data needed to achieve 90% of the full model performance: (a) 40%, (b) 10%, (c) 10%, (d) 6% for LAVENDER (unified as MLM, red) and (a) 60%, (b) 60%, (c) 25%, (d) 10% for task-specific baseline LAVENDER-TS (blue).

- LAVENDER show clearly better generalizability to unseen testing data when trained with limited training data.

Zero-shot Video QA

Method	# pre-train video/images	TGIF			MSRVTT		LSMDC		MSVD
		Act.	Trans.	Frame	MC	QA	MC	FiB	QA
JustAsk [73]	69M / -	-	-	-	-	2.9	-	-	7.5
MERLOT RESERVE [79]	1B/ -	-	-	-	-	5.8	-	31.0	-
BLIP [32]	- / 129M	-	-	-	-	19.2	-	-	35.2
Flamingo [2]	2.1B / 27M	-	-	-	-	19.2	-	-	35.2
FrozenBiLM [74]	- / 10M	-	-	41.9	-	16.9	-	51.5	33.8
All-in-one [62]	283M / -	-	-	-	80.3	-	56.3	-	-
LAVENDER-TS	2.5M / 3M	48.5	47.9	0.0	84.6	0.0	66.9	0.0	0.0
LAVENDER	2.5M / 3M	52.6	**54.1**	16.7	86.7	**4.5**	73.8	34.2	**11.6**
	14M / 16M	**55.1**	53.8	**19.6**	**87.2**	2.7	**73.9**	**36.7**	9.2

Table 4. **Zero-shot Evaluation on Video QA** (top-1 accuracy). Models are evaluated directly after pre-training. BLIP [32] is additionally supervised with VQA v2 [20], and MERLOT RESERVE [79] is pre-trained with additional audio modality and uses GPT-3 [6] to reword questions into masked statements. Flamingo [2] and FrozenBiLM [74] leverage large language models with more than 8x more parameters than the BERT-Base model in LAVENDER.

- LAVENDER can be seamlessly applied to Video QA in a zero-shot manner, with the same MLM head from pre-training
- Compared with previous methods, LAVENDER can achieve competitive ZS performance, even when pre-trained with much less data (5.5M vs. >69M) and without leveraging powerful LLMs

Comparison with SOTA

Method	# Pretrain videos/images	# Params in Backbone	TGIF			MSRVTT		LSMDC		MSVD	Captioning	
			Act.	Trans.	Frame	MC	QA	MC	FiB	QA	MSRVTT	MSVD
ClipBERT [29]	- / 200K	137M	82.8	87.8	60.3	88.2	37.4	-	-	-	-	-
JustAsk [73]	69M / -	166M	-	-	-	-	41.5	-	-	46.3	-	-
MERLOT [80]	180M / -	219M	94.0	96.2	69.5	90.9	43.1	81.7	52.9	-	-	-
VIOLET [15]	183M / 3M	198M	92.5	95.7	68.9	91.9	43.9	82.8	53.7	47.9	-	-
All-in-one [62]	283M / -	110M	95.5	94.7	66.3	92.3	**46.8**	84.4	-	48.3	-	-
SwinBERT [36]	- / -	198M	-	-	-	-	-	-	-	-	53.8	120.6
MV-GPT [54]	53M / -	314M	-	-	-	-	41.7	-	-	-	60.0	-
LAVENDER	2.5M / 3M	198M	**96.6**	**99.1**	72.2	96.6	44.2	86.0	56.9	55.4	58.0	142.9
	14M / 16M		96.3	98.7	**73.5**	**97.4**	45.0	**87.0**	**57.1**	**56.6**	**60.1**	**150.7**

Table 5. Comparison with SOTA on **video QA** (accuracy) and **captioning** (CIDEr).

Method	# Pretrain videos/images	# Params in Backbone	Text-to-Video Retrieval			
			MSRVTT	DiDeMo	MSVD	LSMDC
ClipBERT [29]	- / 200K	137M	22.0 / 46.8 / 59.9	20.4 / 48.0 / 60.8	-	-
Frozen [3]	2.5M / 3.2M	232M	32.5 / 61.5 / 71.2	31.0 / 59.8 / 72.4	45.6 / 79.8 / 88.2	15.0 / 30.8 / 39.8
VIOLET [15]	183M / 3M	198M	34.5 / 63.0 / 73.4	32.6 / 62.8 / 74.7	-	16.1 / 36.6 / 41.2
All-in-one [62]	103M / -	110M	37.9 / 68.1 / 77.1	32.7 / 61.4 / 73.5	-	-
BridgeFormer [19]	- / 400M	~149M	44.9 / 71.9 / 80.3	-	**54.4** / **82.8** / 89.4	21.8 / 41.1 / 50.6
QB-Norm [5]	- / 400M	~149M	47.2 / 73.0 / 83.0	43.3 / 71.4 / 80.8	47.6 / 77.6 / 86.1	22.4 / 40.1 / 49.5
CAMoE [11]	- / 400M	~149M	**47.3** / **74.2** / **84.5**	43.8 / 71.4 / 79.9	49.8 / 79.2 / 87.0	25.9 / 46.1 / 53.7
LAVENDER	2.5M / 3M	198M	37.8 / 63.8 / 75.0	47.4 / 74.7 / 82.4	46.3 / 76.9 / 86.0	22.2 / 43.8 / 53.5
	14M / 16M		40.7 / 66.9 / 77.6	**53.4** / **78.6** / **85.3**	50.1 / 79.6 / 87.2	**26.1** / **46.4** / **57.3**

Table 6. Comparison with SOTA on **text-to-video-retrieval** (R1/5/10). CAMoE [11] assumes the model can see all queries during testing.

- Without any task-specific architectures, LAVENDER outperforms the prior state-of-the-art on 11 out of 14 benchmarks considered

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu,
Ce Liu, Lijuan Wang

Paper Tag: THU-PM-240