JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

MELTR: Meta Loss Transformer for Learning to Fine-Tune Video Foundation Models

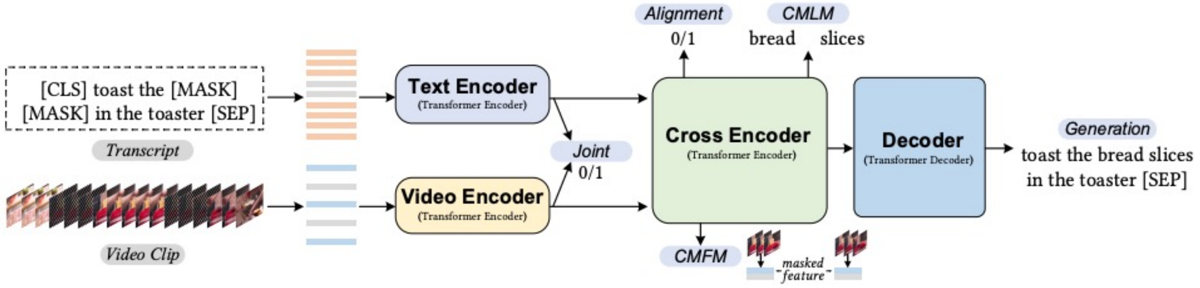
Dohwan Ko[1*], Joonmyung Choi[1*], Hyeong Kyu Choi[1], Kyoung-Woon On[2], Byungseok Roh[2], Hyunwoo J. Kim[1]

[1]Department of Computer Science and Engineering, Korea University [2]Kakao Brain

Poster Tag: THU-AM-344

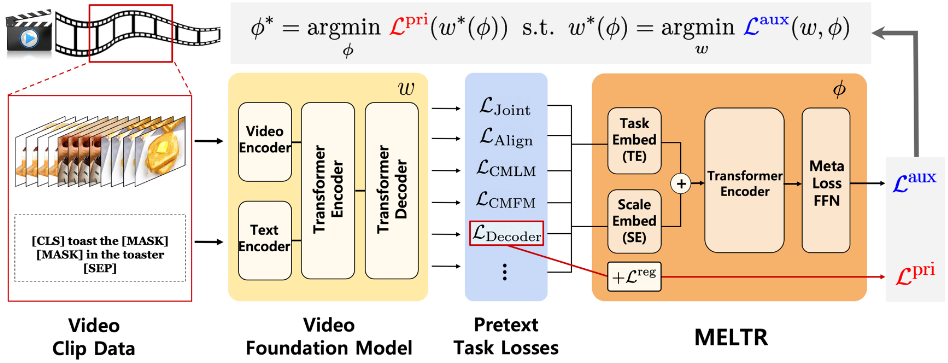
Preview

Motivation



$$\mathcal{L}_{\text{UniVL}} = \mathcal{L}_{\text{Joint}} + \mathcal{L}_{\text{CMLM}} + \mathcal{L}_{\text{CMFM}} + \mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{Decoder}}$$

Method



Idea

- How can *automatically* combine various pretext task loss functions to assist learning of the target task?
- Use a meta-learning-based *auxiliary learning* framework.

Results

- MELTR significantly outperforms the baselines across **three backbone models** on **five datasets**.

Models	MSRVTT-7k R@1↑	MSRVTT-7k R@5↑	MSRVTT-7k R@10↑	MSRVTT-7k MedR↓	MSRVTT-9k R@1↑	MSRVTT-9k R@5↑	MSRVTT-9k R@10↑	MSRVTT-9k MedR↓
MIL-NCE [52]	9.9	24.0	32.4	29.5	-	-	-	-
JSFusion [55]	10.2	31.2	43.2	13	-	-	-	-
HowTo100M [35]	14.9	40.2	52.8	9	-	-	-	-
HERO [26]	16.8	43.4	57.7	-	-	-	-	-
ClipBERT [56]	22.2	46.8	59.9	6	-	-	-	-
MMT [20]	-	-	-	-	26.6	57.1	69.6	4
T2VLAD [57]	-	-	-	-	29.5	59.0	70.1	4
TACo [24]	19.2	44.7	57.2	7	28.4	57.8	71.2	4
VideoCLIP [54]	-	-	-	-	30.9	55.4	66.8	-
Frozen [58]	-	-	-	-	32.5	61.5	71.2	3
UniVL-Joint [7]	20.6	49.1	62.9	6	27.2	**55.7**	**68.7**	**4**
UniVL-Align [7]	21.2	49.6	63.1	6	-	-	-	-
UniVL + **MELTR**	**28.5**	**55.5**	**67.6**	**4**	**31.1**	**55.7**	68.3	**4**
Violet [16]	31.7	60.1	74.6	**3**	34.5	63.0	73.4	-
Violet + **MELTR**	**33.6**	**63.7**	**77.8**	**3**	**35.5**	**67.2**	**78.4**	**3**
All-in-one [17]	34.4	65.4	75.8	-	37.9	68.1	77.1	-
All-in-one + **MELTR**	**38.6**	**74.4**	**84.7**	-	**41.3**	**73.5**	**82.5**	-

Huaishao Luo et al. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arxiv, 2021.

Motivation

Video Foundation Models

- Large-scale foundation models pretrained on huge amounts of data.

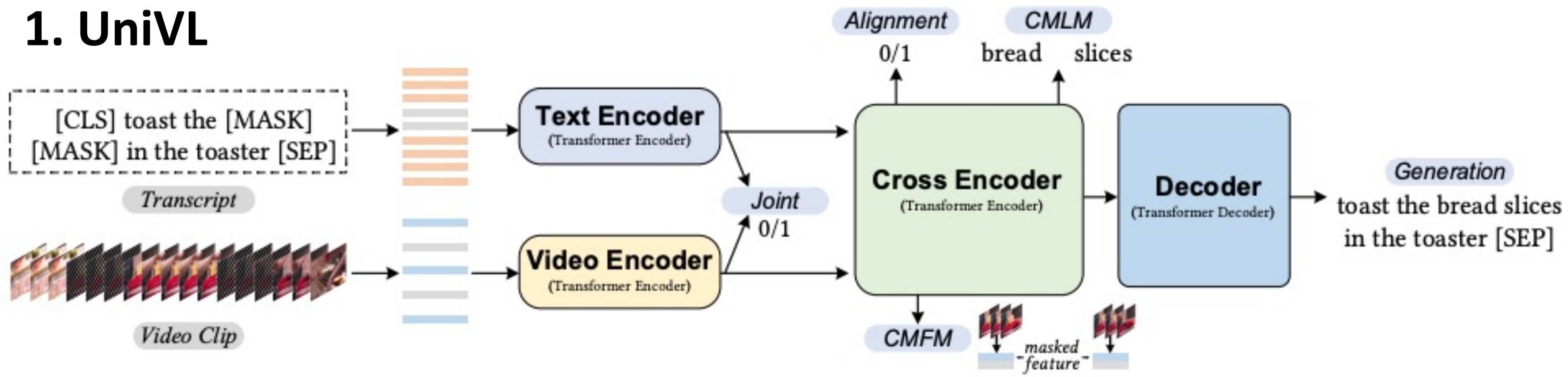
- Advantages of adaptability and generalizability to a wide range of downstream tasks.

- Pre-trained with a *linear* combination of various pretext tasks.
  Ex) text-video alignment (VTC, VTM), MLM, MFM, and generation.

Motivation

Video Foundation Models

1. UniVL



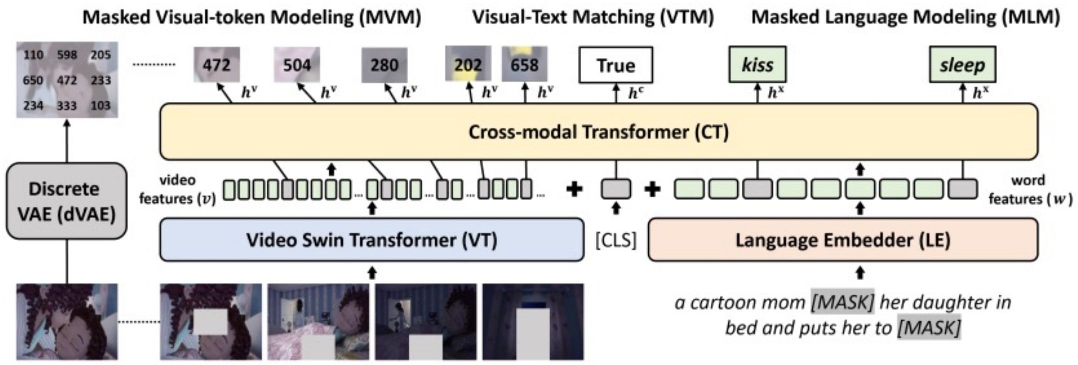
$$\boxed{\mathcal{L}_{\mathrm{UniVL}} =}\boxed{\mathcal{L}_{\mathrm{Joint}}}\boxed{+ \mathcal{L}_{\mathrm{CMLM}} + \mathcal{L}_{\mathrm{CMFM}} + \mathcal{L}_{\mathrm{Align}} +}\boxed{\mathcal{L}_{\mathrm{Decoder}}}$$

Huaishao Luo et al. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arxiv, 2021.

Motivation

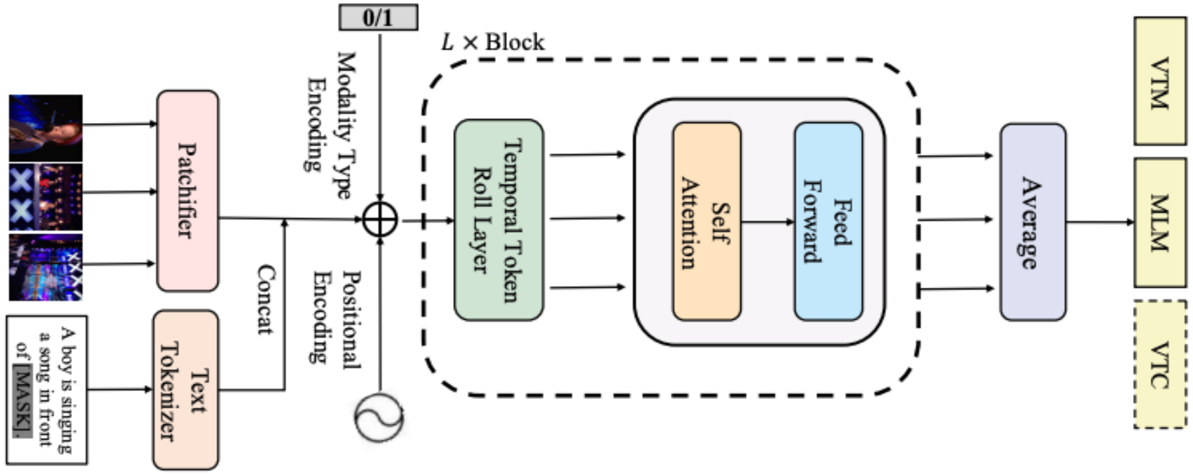
Video Foundation Models

2. Violet



$$\mathcal{L}_{\text{Violet}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MVM}}$$

3. All-in-one



$$\mathcal{L}_{\text{All-in-one}} = \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{VTC}}$$

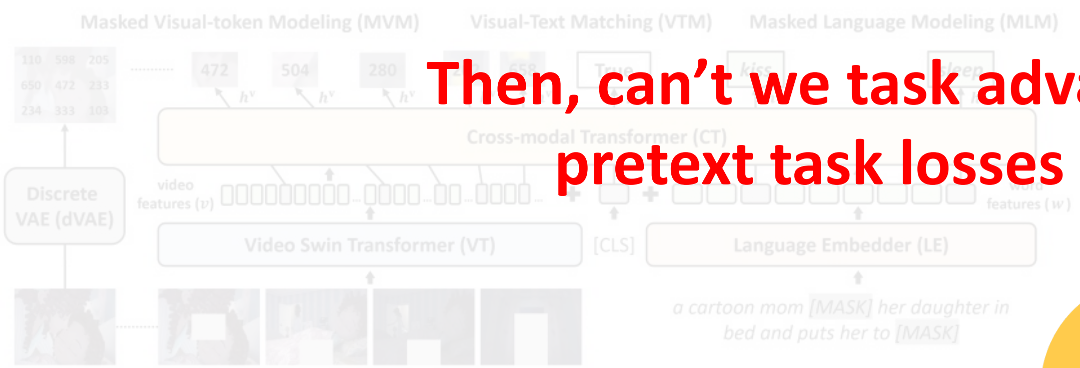
Tsu-Jui Fu et al. Violet: End-to-end video-language transformers with masked visual-token modeling. arxiv, 2023.

Jinpeng Wang et al. All in One: Exploring Unified Video-Language Pre-training. CVPR, 2023.

Motivation

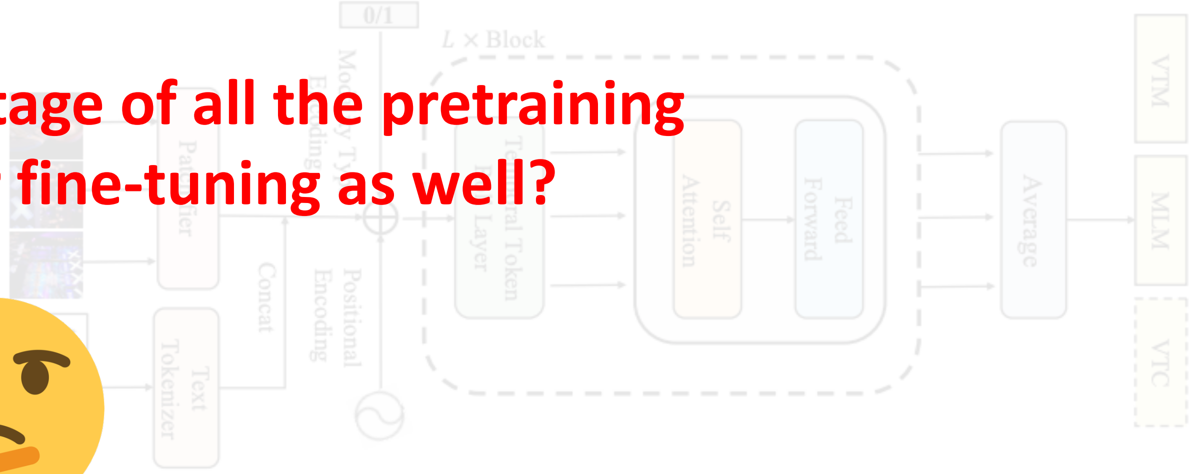
Video Foundation Models

2. Violet



$$\mathcal{L}_{\text{Violet}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MVM}}$$

3. All-in-one



$$\mathcal{L}_{\text{All-in-one}} = \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{VTC}}$$

Then, can't we task advantage of all the pretraining pretext task losses for fine-tuning as well?

Tsu-Jui Fu et al. Violet: End-to-end video-language transformers with masked visual-token modeling. arxiv, 2023.
Jinpeng Wang et al. All in One: Exploring Unified Video-Language Pre-training. CVPR, 2023.

Idea

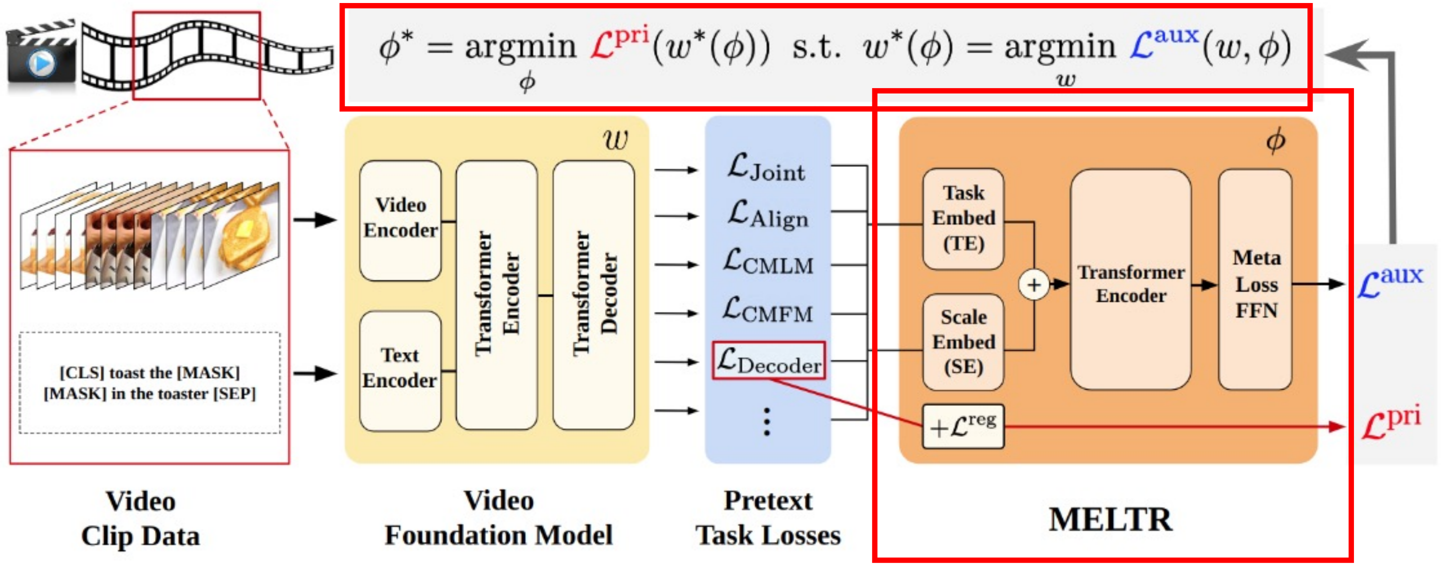
Auxiliary Learning

	Multi-task learning	Auxiliary learning
Use multiple tasks?	Yes	Yes
Purpose	Aims for generalization across tasks	Focues only on the primary task by taking advantage of auxiliary tasks

- **Learns to adaptively leverage multiple auxiliary tasks** to assist learning of the primary task (based on Meta-learning).
- The pretext task losses can be **unified into a single auxiliary loss** to be optimized in a way that helps the target downstream task.

Method

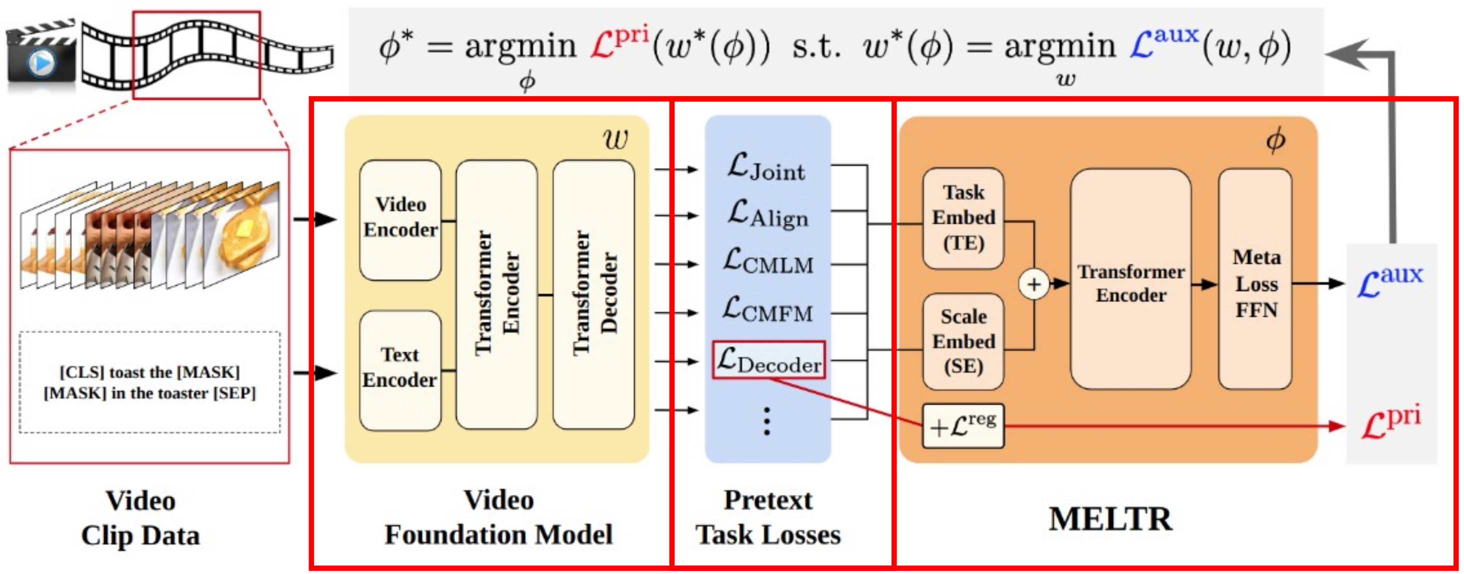
Meta Loss Transformer (MELTR)



- A plug-in module for meta auxiliary learning
- Adopt Transformer architecture.
- MELTR is optimized to help learning of the primary task.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Meta Loss Transformer (MELTR)



$$\mathcal{L}^{\text{pri}} = \mathcal{L}_0 + \gamma \mathcal{L}^{\text{reg}}, \quad \mathcal{L}^{\text{aux}} = \text{MELTR}(\boldsymbol{\ell}; \phi)$$

- Calculate losses:

$$\ell_t = \mathcal{L}_t(\mathcal{F}(x; w), y_t)$$

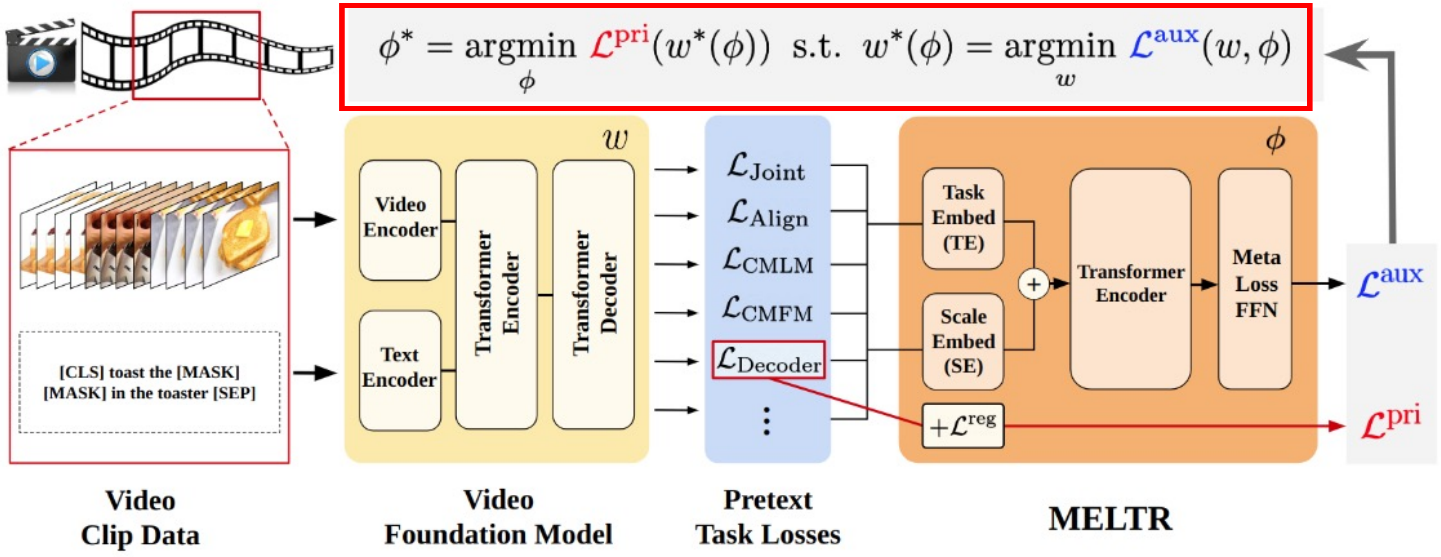
- Transform losses into a single unified loss:

$$\mathcal{L}^{\text{aux}} := \text{MELTR}(\boldsymbol{\ell}; \phi)$$

- Regularization term:

$$\mathcal{L}^{\text{reg}} = \left| \text{MELTR}(\boldsymbol{\ell}; \phi) - \sum_{t=0}^{T} \ell_t \right|$$

Meta Loss Transformer (MELTR)



$$\mathcal{L}^{\text{pri}} = \mathcal{L}_0 + \gamma\mathcal{L}^{\text{reg}}, \quad \mathcal{L}^{\text{aux}} = \text{MELTR}(\boldsymbol{\ell};\phi)$$

- Objective function

$$\phi^* = \arg\min_{\phi}\ \mathcal{L}^{\text{pri}}(w^*(\phi))$$

$$\text{s.t. } w^*(\phi) = \arg\min_{w}\ \mathcal{L}^{\text{aux}}(w,\phi)$$

- For K steps, update backbone foundation model:

$$w^{(k+1)} = w^{(k)} - \alpha \cdot \nabla_w \mathcal{L}^{\text{aux}}$$

- Then, update MELTR:

$$\begin{aligned}\phi^* &= \phi - \beta \cdot \nabla_\phi \mathcal{L}^{\text{pri}}(w^{(K)}(\phi)) \\ &= \phi + \beta \cdot \left(\nabla_w \mathcal{L}^{\text{pri}} \cdot \nabla_\phi \nabla_w \mathcal{L}^{\text{aux}}\right)\end{aligned}$$

Experiments

Quantitative results

Models	TGIF-QA Action	TGIF-QA Transition	TGIF-QA Frame	MSVD-QA
HME [61]	73.9	77.8	53.8	33.7
HCRN [62]	75.0	81.4	55.9	36.1
QueST [63]	75.9	81.0	59.7	36.1
ClipBERT [64]	82.9	87.5	59.4	-
Violet [16]	92.5	95.7	62.3†	47.9
Violet + **MELTR**	**95.4**	**97.5**	**63.4**	**51.7**

Video question answering

Models	BA↑	F1↑	MAE↓	Corr↑
MulT	83.0	82.8	0.870	0.698
FMT	83.5	83.5	0.837	0.744
UniVL	84.6	84.6	0.781	0.767
UniVL + **MELTR**	**85.3**	**85.4**	**0.759**	**0.789**

Multi-modal sentiment analysis on CMU-MOSI

Models	R@1↑	R@5↑	R@10↑	MedR↓
HGLMM-FV-CCA [55]	4.6	21.6	14.3	75
HowTo100M [35]	8.2	35.3	24.5	24
ActBERT [29]	9.6	26.7	38.0	19
MIL-NCE [56]	15.1	38.0	51.2	10
COOT [57]	16.7	40.2	25.3	9
TACo [24]	29.6	59.7	72.7	9
VideoCLIP [58]	32.2	62.6	**75.0**	-
UniVL-Joint [7]	22.2	52.2	66.2	5
UniVL-Align [7]	28.9	57.6	70.0	4
UniVL + **MELTR**$^-$	33.4	62.5	73.3	**3**
UniVL + **MELTR**	**33.7**	**63.1**	74.8	**3**

Text-to-video retrieval on YouCook2

Models	Modality	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
EMT [65]	V	7.53	4.38	11.55	27.44	0.38
CBT [27]	V	-	5.12	12.97	30.44	0.64
ActBERT [29]	V	8.66	5.41	13.30	30.56	0.65
VideoBERT [28]	V	6.33	3.81	10.81	27.14	0.47
COOT [57]	V	17.97	11.30	19.85	37.94	0.57
VideoBERT [28]	V+T	7.59	4.33	11.94	28.80	0.55
DPC [66]	V+T	7.60	2.76	18.08	-	-
AT+Video [67]	V+T	-	9.01	17.77	36.65	1.12
UniVL [7]	V	16.46	11.17	17.57	40.09	1.27
UniVL + **MELTR**	V	17.35	11.98	18.19	41.28	1.38
UniVL [7]	V+T	23.87	17.35	22.35	46.52	1.81
UniVL + **MELTR**	V+T	**24.12**	**17.92**	**22.56**	**47.04**	**1.90**

Video captioning on YouCook2

Models	Modality	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
PickNet [68]	V	-	35.6	26.8	58.2	41.0
PickNet [68]	V+T	-	38.9	27.2	59.5	42.1
MARN [69]	V	-	40.4	28.1	60.7	47.1
SibNet [70]	V	-	40.9	27.5	60.2	47.5
OA-BTG [71]	V	-	41.4	28.2	-	46.9
POS-VCT [72]	V	-	42.3	**29.7**	**62.8**	49.1
ORG-TRL [73]	V	-	43.6	28.8	62.1	50.9
UniVL* [7]	V	53.42	41.79	28.94	60.78	50.04
UniVL + **MELTR**	V	**55.88**	**44.17**	29.26	62.35	**52.77**

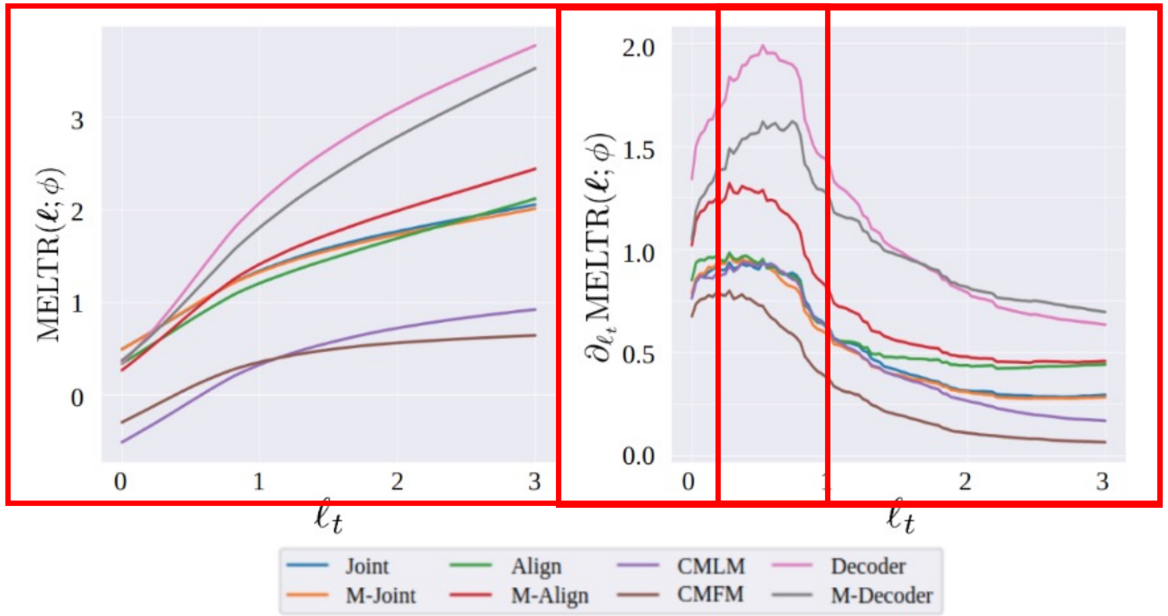
Video captioning on MSRVTT

Models	R@1↑	R@5↑	R@10↑	MedR↓
MIL-NCE [56]	9.9	24.0	32.4	29.5
JSFusion [59]	10.2	31.2	43.2	13
HowTo100M [35]	14.9	40.2	52.8	9
HERO [26]	16.8	43.4	57.7	-
ClipBERT [60]	22.2	46.8	59.9	6
TACo [24]	19.2	44.7	57.2	7
UniVL-Joint [7]	20.6	49.1	62.9	6
UniVL-Align [7]	21.2	49.6	63.1	6
UniVL + **MELTR**	**28.5**	**55.5**	**67.6**	**4**
Violet [16]	31.7†	60.1†	74.6†	3†
Violet + **MELTR**	**33.6**	**63.7**	**77.8**	**3**
All-in-one [17]	34.4	65.4	75.8	-
All-in-one + **MELTR**	**38.6**	**74.4**	**84.7**	-

Text-to-video retrieval on MSRVTT

Experiments

Analysis: Non-linear loss transformation



$$\partial_{\ell_t}\mathrm{MELTR}(\boldsymbol{\ell};\phi) := \frac{\partial}{\partial \ell_t}\mathrm{MELTR}(\boldsymbol{\ell};\phi)$$

- Non-linearly correlated.
- $\partial_{\ell_t}\mathrm{MELTR}(\boldsymbol{\ell};\phi)$ have relatively higher values around $\ell_t = 0.5$.
  - Focus on reasonably challenging samples
- MELTR is more sensitive to $\mathcal{L}_{\text{Decoder}}$ and $\mathcal{L}_{\text{M-Decoder}}$ than $\mathcal{L}_{\text{CMFM}}$.

Experiments

Analysis: Adaptive task re-weighting

Models	Coefficient of each task								Video captioning on YouCook2				
	$\mathcal{L}_{\text{Joint}}$	$\mathcal{L}_{\text{M-Joint}}$	$\mathcal{L}_{\text{Align}}$	$\mathcal{L}_{\text{M-Align}}$	$\mathcal{L}_{\text{CMLM}}$	$\mathcal{L}_{\text{CMFM}}$	$\mathcal{L}_{\text{Decoder}}$	$\mathcal{L}_{\text{M-Decoder}}$	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
(A)	0	0	0	0	0	0	1	0	22.79	16.54	21.73	45.85	1.78
(B)	0	0	0	0	0	0	1	1	23.42	17.14	22.27	46.65	1.85
(C)	1	1	1	1	1	1	1	1	21.72	15.93	20.89	45.16	1.79
(D)	1	1	1	1	1	0	1	1	21.99	16.10	21.09	45.35	1.85
(E)	1	1	1	1	1	0	8	8	23.31	17.23	21.98	46.26	1.85
MELTR	ADAPTIVE								**24.12**	**17.92**	**22.56**	**47.04**	**1.90**

Conclusion

- MELTR **learns to integrate various pretext task losses into one loss function** to boost the performance of the target downstream task.
- By plugging MELTR into various foundation models, our method **outperformed video foundation models as well as task-specific models** on a wide range of downstream tasks.
- We provide in-depth qualitative analyses of how MELTR adequately **transforms** individual loss functions and **melts** them into an effective unified loss function.