Weakly Supervised Video Representation Learning with Unaligned Text for Sequential Videos

CVPR 2023

Sixun Dong[1*] Huazhang Hu[1*]

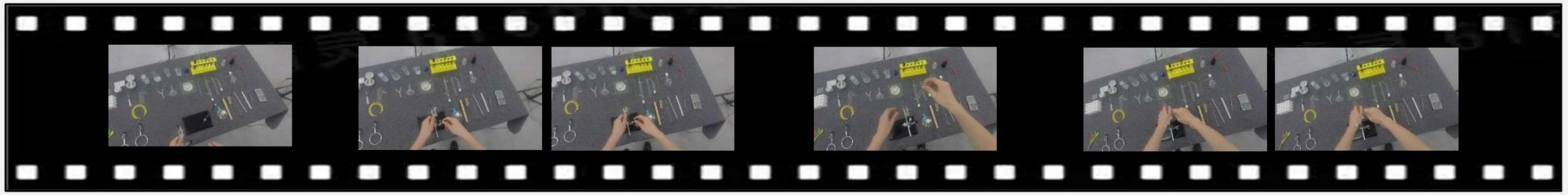
Dongze Lian [1,2], Weixin Luo [3], Yicheng Qian [1], Shenghua Gao [1,4,5†]

*Equal Contributions †Corresponding authors

[1]ShanghaiTech University [2]National University of Singapore [3]Meituan

[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging

[4]Shanghai Engineering Research Center of Energy Efficient and Custom AI IC



Texts : [Take up the iron clamp, Fix on the iron stand, Take up the test tube, Screw the iron clamp]

上海科技大学 ShanghaiTech University

SVIP LAB ShanghaiTech Vision and Intelligent Perception

NUS National University of Singapore

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

(1) Motivation

➢ Sequential videos

Texts

- take up the jar
- uncover the jar cap
- pour the jar
- cover the jar with the jar cap
- put down the jar

Texts

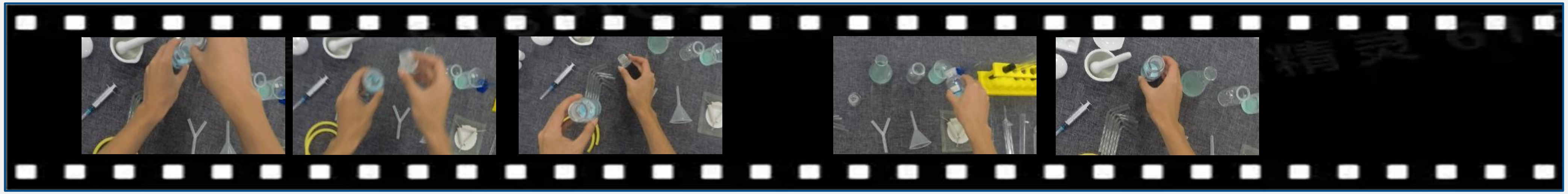
- take up the jar
- uncover the jar cap
- <u>put down the jar cap</u>
- <u>pour the jar</u>
- put down the jar

[1] The examples are from CSV dataset. “SVIP: Sequence VerIfication for Procedures in Videos". In CVPR 2022

(1) Motivation

➤ Sequential Video

Texts : [take up the jar, uncover the jar cap, pour the jar, cover the jar with the jar cap, put down the jar]



Texts : [take up the jar, uncover the jar cap, put down the jar cap, pour the jar, put down the jar]

- No time-stamp annotation
- Step annotations
- Multiple sequential actions
- Similar ordering of actions

[1] The examples are from CSV dataset. “SVIP: Sequence VerIfication for Procedures in Videos". In CVPR 2022

(2) Previous Works

➢ Sequence verification for procedures in videos

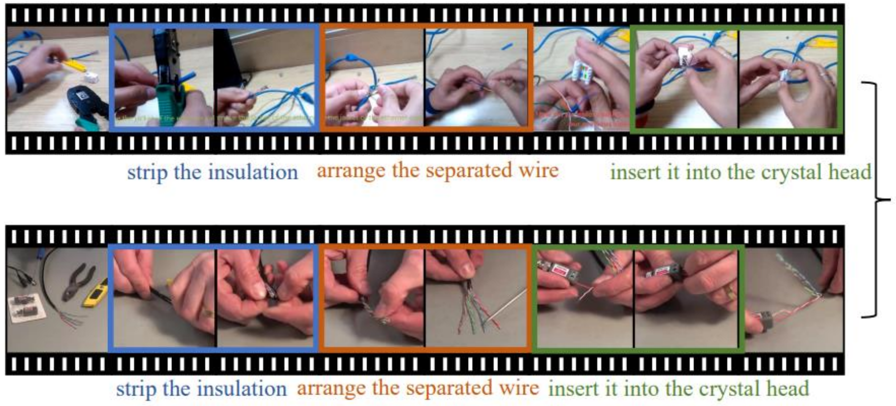


Figure 2. Positive video pair (Yicheng Qian et at.)

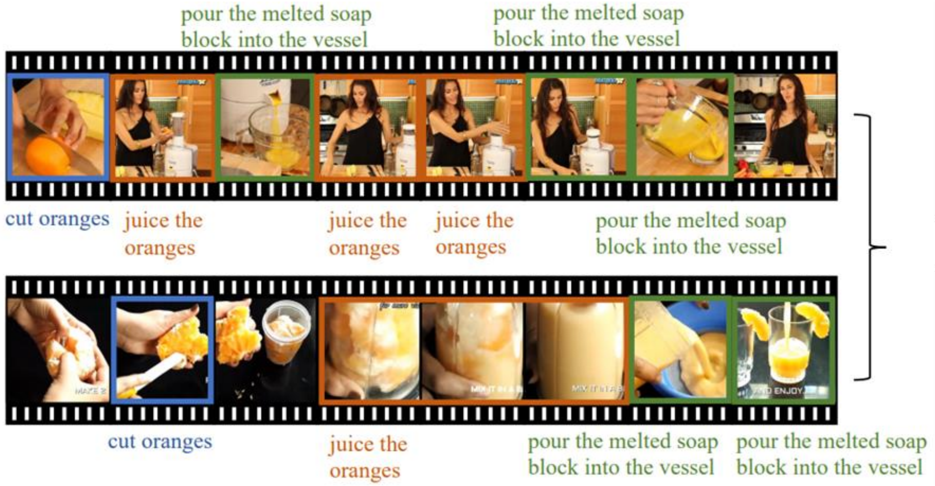


Figure 3. Negative video pair (Yicheng Qian et at.)

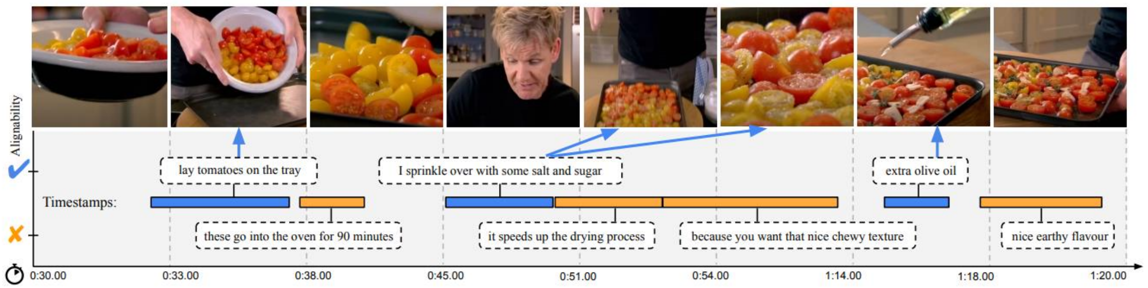
- Slightly different step
- Rely on additional class information
- Under supervision

✓ Measure video representation

✓ Focus on every procedure

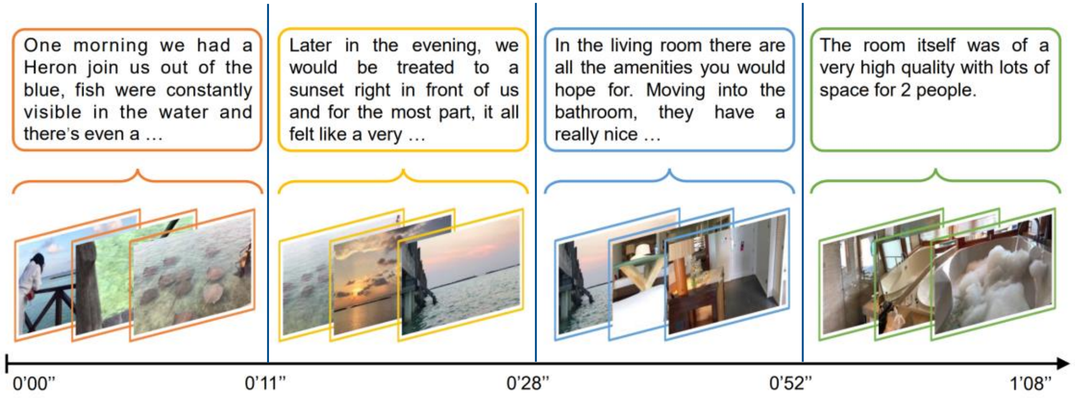
[1] Yicheng Qian et at., " SVIP: Sequence VerIfication for Procedures in Videos." In CVPR, 2022.

(2) Previous Works



(1) Visual-textual mis-alignment (Han Tenda et al., CVPR 2022 Oral)[2]

- ➢ Visual-textual mis-alignment
  - ◻ Noisy time-step annotations
  - ◻ Ignore missing fine-grained alignment



(2) Video-paragraph pair (Yuchong Sun et al., NeurIPS 2022)[3]

- ➢ Video-paragraph pair
  - ◻ Segmented time-step annotations
  - ◻ Not fine-grained enough

---

[2] Han Tenda et al., "Temporal alignment networks for long-term video." In CVPR, 2022.
[3] Yuchong Sun et al., "Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning". In NeurIPS 2022

(3) Method

- ✓ Propose a contrastive learning framework
- ✓ Design multiple granularity contrastive loss

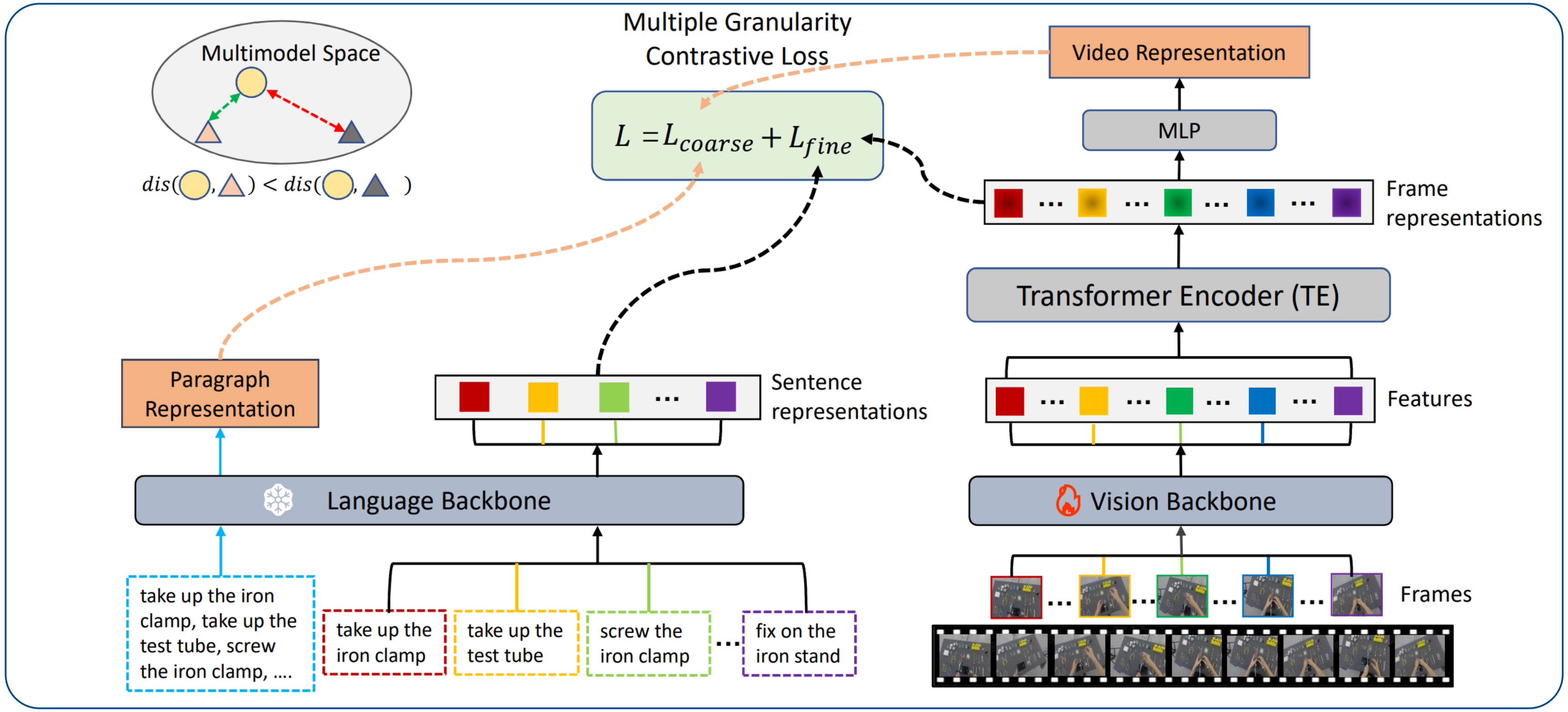


Figure 5. Overview of our framework.

(3) Method

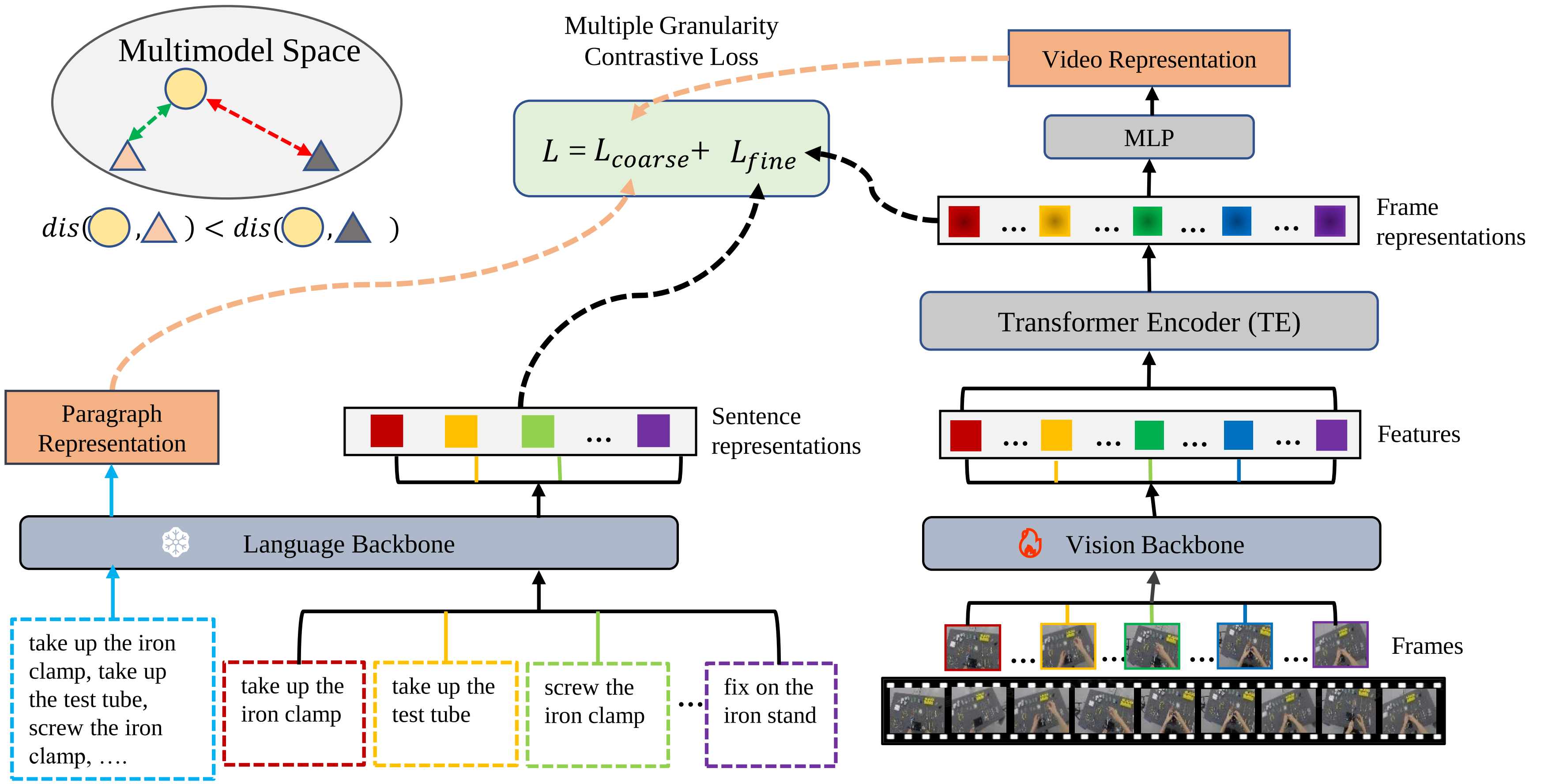
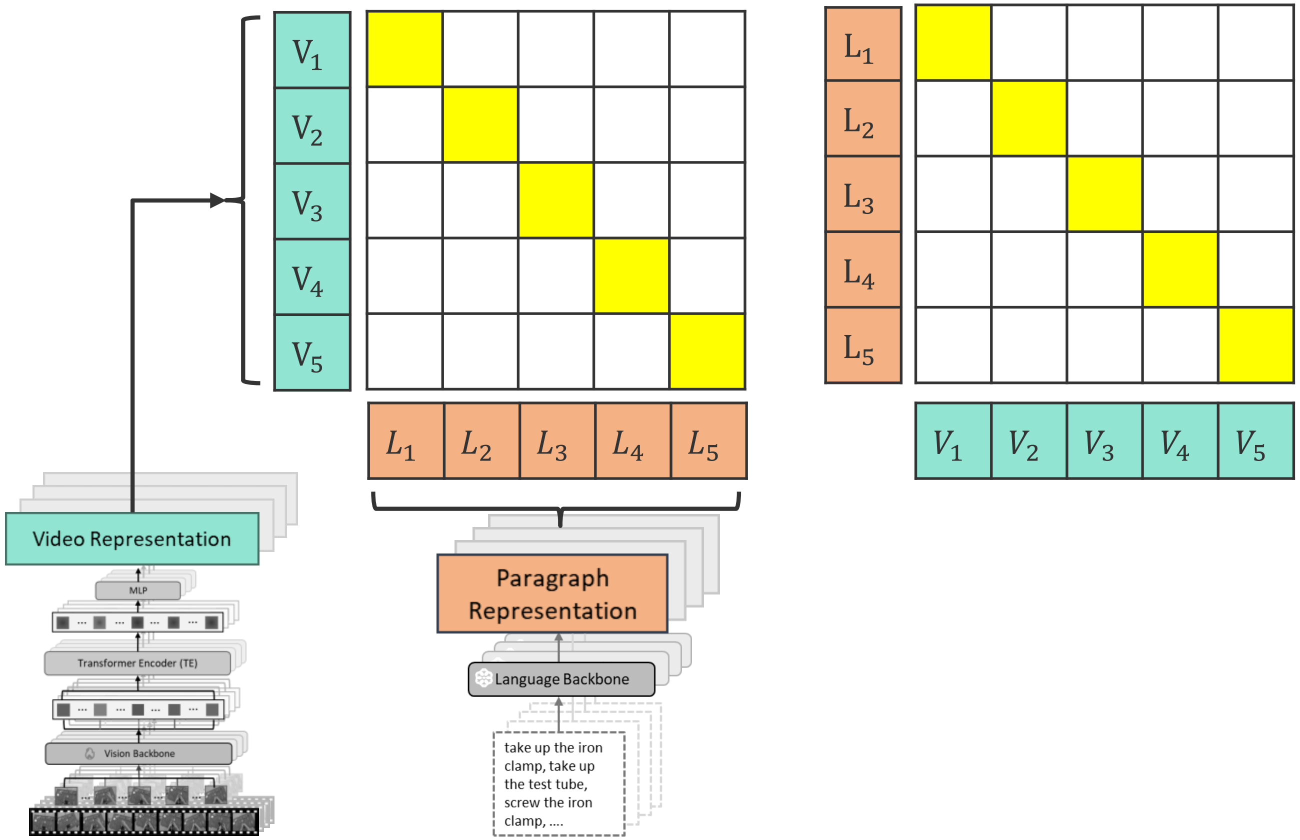


Figure 5. Overview of our framework.

(3) Method

➢ Coarse-grained Loss



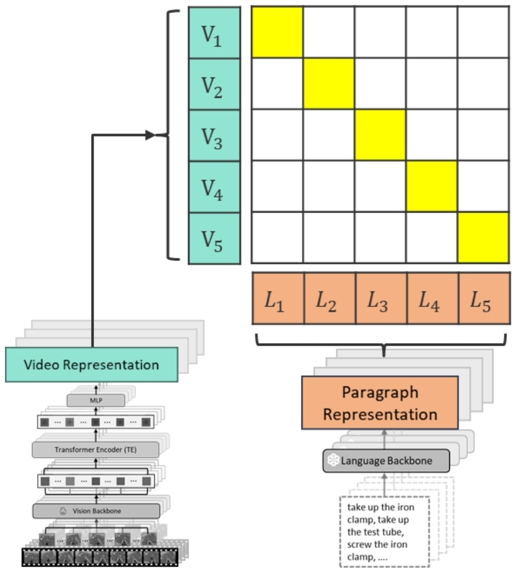
$$L_{\text{InfoNCE}}(V, L) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\varphi(v_i, l_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\varphi(v_j, l_j)/\tau\right)}$$

$$\varphi(v_i, l_i) = \frac{v_i}{\|v_i\|}\cdot\frac{l_i^T}{\|l^T\|}$$

$$L_{\text{coarse}} = L_{\text{InfoNCE}}(V, L) + L_{\text{InfoNCE}}(L, V)$$

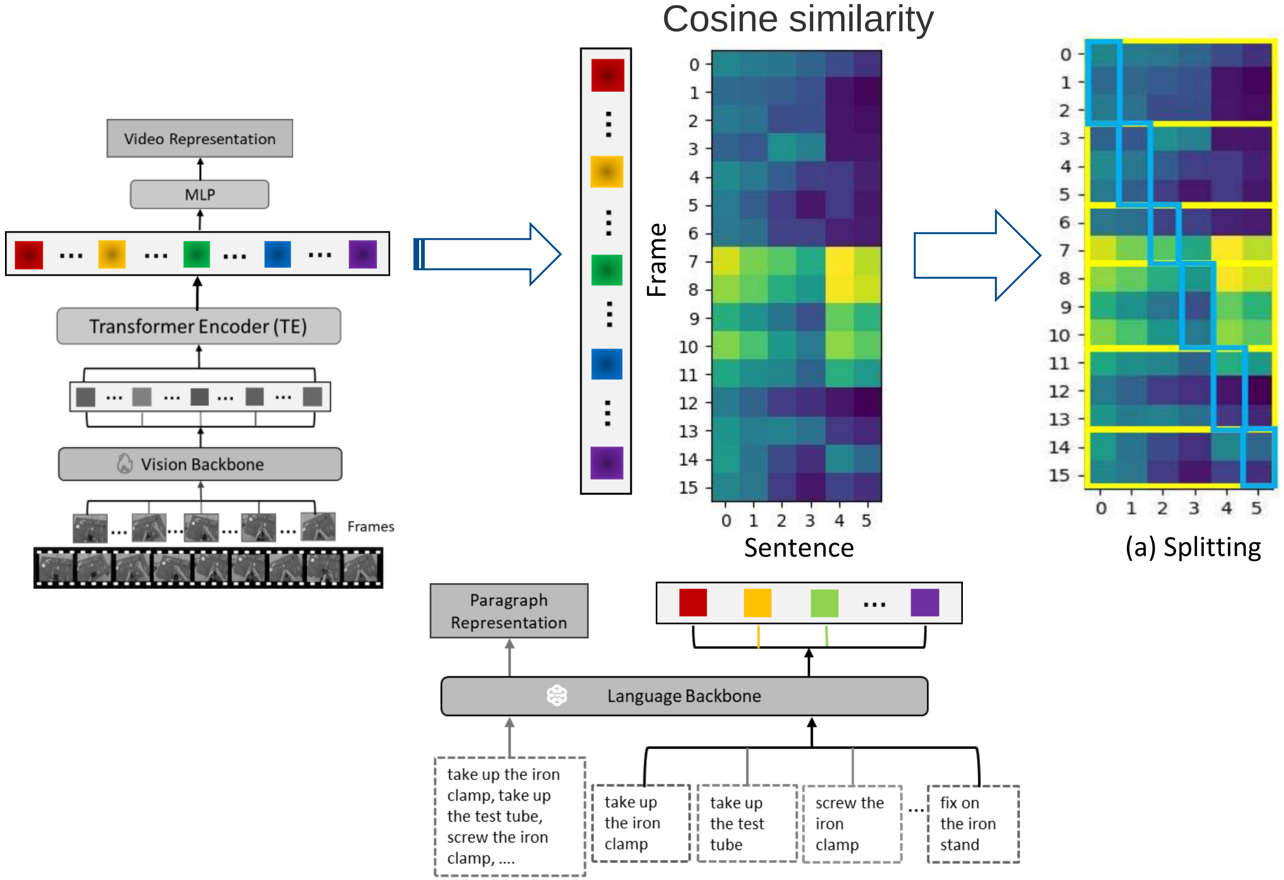
(3) Method

➢ Coarse-grained loss



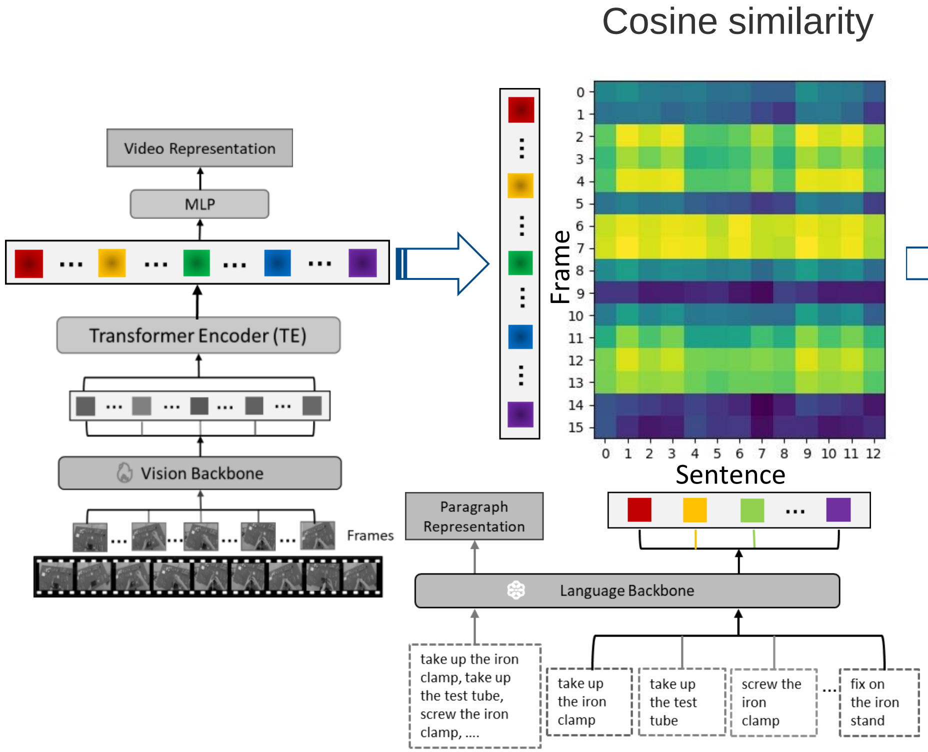
$$L_{\text{InfoNCE}}(V, L) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\varphi(v_i, l_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\varphi(v_j, l_j)/\tau\right)}$$

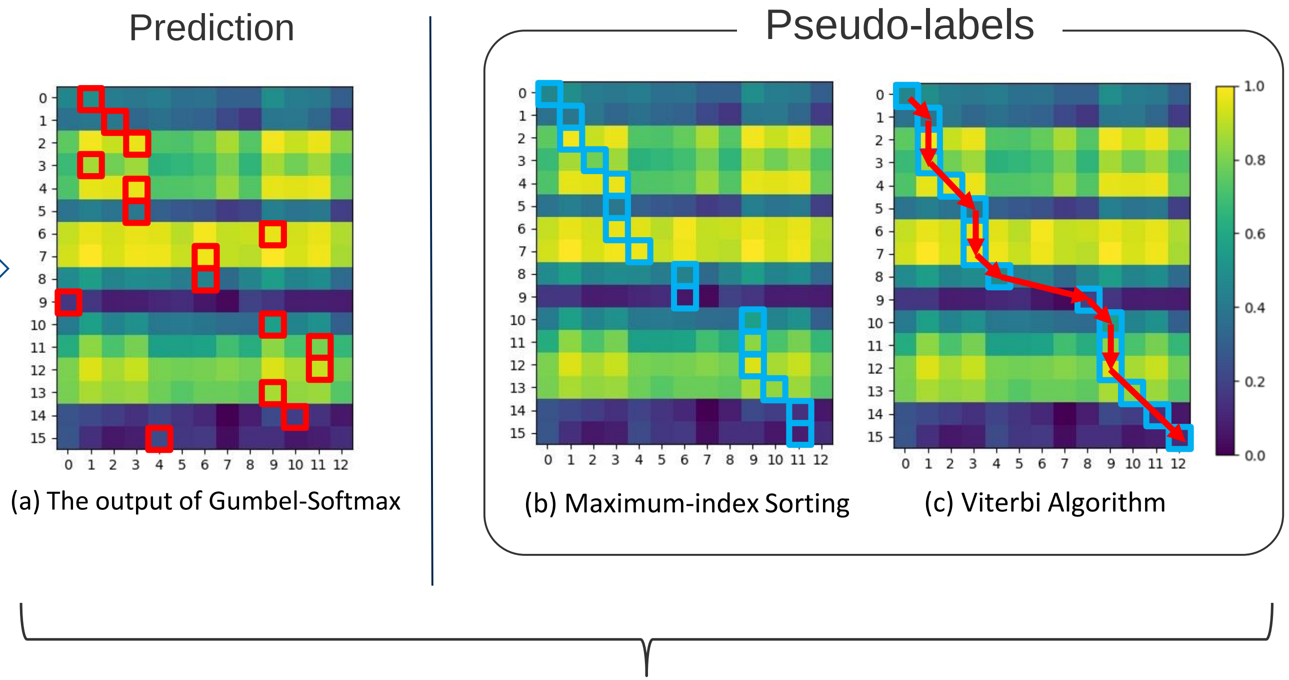
➢ Fine-grained loss



(3) Method

➢ Fine-grained contrastive loss





$$L_{fine} = CE(\psi_{\text{preds}}(H, S), \phi_{\text{pseudo}}(H, S)) + CE(\psi_{\text{preds}}(S, H), \phi_{\text{pseudo}}(S, H))$$

(4) Experiments

➢ Evaluation matrices

$$d = dis(v_1, v_2)$$

$$y = \begin{cases} 1, d \leq \tau \\ 0, otherwise \end{cases}$$

➢ Video sequence verification

Method	Text Encoder	Weakly Supervised (w/o CLS)		
		CSV	Diving-SV	COIN-SV
MIL-NCE [30]	MLP [30]	53.02	58.49	47.95
CAT [35]	CLIP [37]	70.63	77.87	47.70
VideoSwin [28]+MLP	CLIP [37]	62.48	60.88	**54.73**
CLIP [37]+TE [10]+Pool	CLIP [37]	58.67	72.13	49.79
CLIP [37]+TE [10]+MLP	CLIP [37]	74.82	81.47	50.13
Ours	CLIP [37]	**79.80**	**85.19**	52.56

Table 1. Results of representation learning for weakly supervised video sequence verification task.

(4) Experiments

- Video sequence verification

Method	Pre-train	Supervised (w CLS)		
		CSV	Diving-SV	COIN-SV
MIL-NCE [29]	HowTo100M [30]	56.16	63.43	47.80
Swin [26]	K-400 [5]	54.06	73.10	43.70
TRN [57]	K-400 [5]	80.32	80.69	57.19
CAT [34]	K-400 [5]	83.02	83.11	51.13
CLIP [36]+TE [10]+MLP	CLIP [36]	79.38	83.48	48.50
Ours (weakly supervised)	CLIP [36]	79.80	85.19	52.56
Ours	CLIP [36]	**86.92**	**86.09**	**59.57**

Method	Backbone	Weakly supervised (w/o CLS)			Supervised (w CLS)		
		Def.	No Rep.	Rep.	Def.	No Rep.	Rep.
CAT [1]	ResNet50	47.70	57.82	49.99	51.13	63.25	45.96
CLIP+TE+MLP	CLIP-ViT	50.83	65.28	53.73	48.50	65.21	51.25
Ours	CLIP-ViT	**52.55**	**68.98**	**56.16**	**59.57**	**77.78**	**54.95**

Table 2. Results of downstream video sequence verification task under supervised.

(4) Experiments

- Text-to-video matching

Method	Text-to-Video Matching
	CSV-Matching
MIL-NCE [29]	60.02
CAT [34]	53.54
CLIP [36] +TE [10] +MLP	62.67
Ours	**65.23**

Table 3. Results of text-to-video matching task on our proposed benchmark *CSV-Matching*. We evaluate the results using AUC.

- Video classification

Method	Backbone	Loss	Classification(Acc)
CAT [40]	ResNet-50	CLS, SEQ	61.08
CLIP [42]+TE+MLP	CLIP-ViT	CLS, SEQ	63.24
Ours (w/o multi-grained loss)	CLIP-ViT	CLS, SEQ	-
Ours(CLS)	CLIP-ViT	CLS, SEQ, Multi-grained loss	69.57

Table 4. Results of video classification on CSV.

(5) Ablation Studies

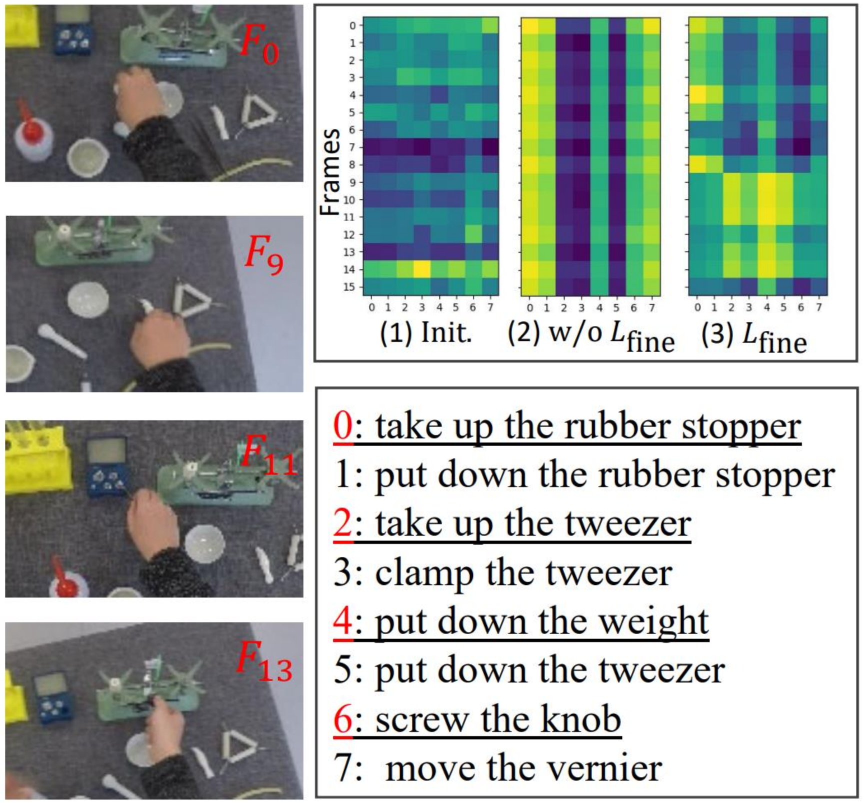


Figure 4. **Visualization** of ablation study about fine-grained contrastive loss.

Method	L_{fine}	L_{coarse}	CSV
Ours (w CLS)	✗	✗	83.58
	✓	✗	84.85
	✗	✓	84.32
	✓	✓	**86.92**

Table 6. Ablation studies of our proposed multiple granularity contrastive loss on CSV. To verify the effectiveness of L_{fine} and L_{coarse} separately, we conduct experiments on video verification task.

Method	L_{fine}	Pseudo-label generation	CSV
Ours	✗	✗	74.82
	✓	split	72.75
		viterbi	78.46
		sort	**79.80**

Table 7. Ablation studies of the type of pseudo-label generation on our proposed method.

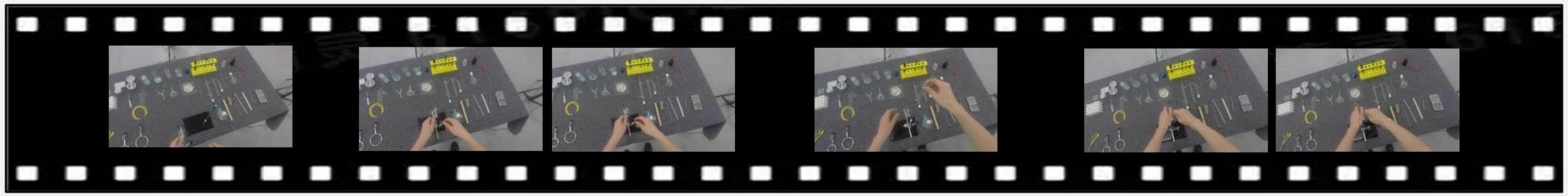
Weakly Supervised Video Representation Learning with Unaligned Text for Sequential Videos

(CVPR 2023)

Sixun Dong* Huazhang Hu *

Dongze Lian， Weixin Luo, Yicheng Qian, Shenghua Gao †

*Equal Contributions †Corresponding authors



Paper: https://arxiv.org/abs/2303.12370

Code: https://github.com/svip-lab/WeakSVR

Code

Paper