Vita-clip: Video and text adaptive clip via multimodal prompting

[1]Syed Talal Wasim [1]Muzammal Naseer [1,2]Salman Khan [1,3]Fahad Shahbaz Khan [4]Mubarak Shah
[1]Mohamed Bin Zayed University of AI, [2]Australian National University, [3]Linköping University, [4]University of Central Florida

CVPR 2023 ID: THU-PM-232

Presenter: Syed Talal Wasim

Research Assistant,
Department of Computer Vision,
Mohamed Bin Zayed University of AI (MBZUAI)

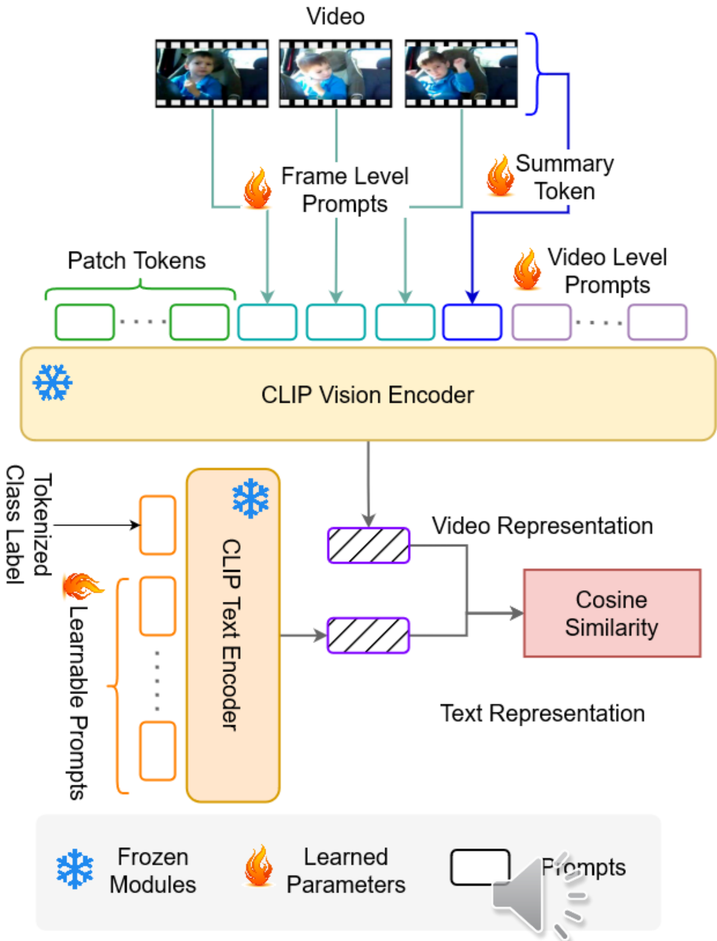
Challenges

- Trade-off in adopting Image Language models to Videos
  - Finetuning the backbone reduces zeroshot performance
  - Frozen backbone results in poor supervised performance

- State-of-the-art methods such as XCLIP tend to have separate training schemes for supervised and zeroshot settings
  - Essentially two different models!!!
  - Might as well have separate models for supervised/zeroshot

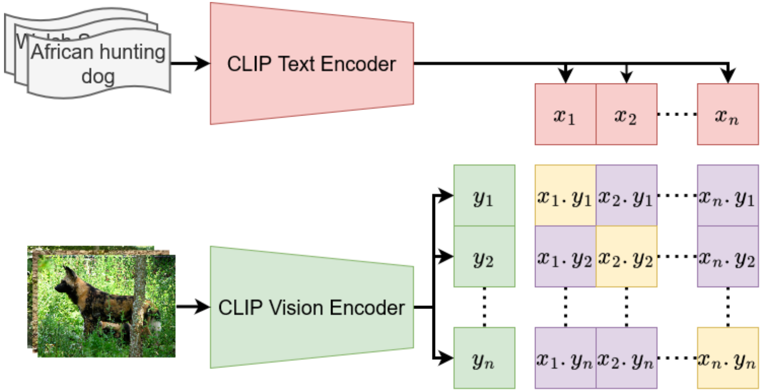
Solution Overview

- Freeze CLIP backbone
  - Retain pretrained generalization
- Introduce prompts on vision and text encoders
  - Frame-level prompts to model per-frame information
  - Summary prompt to summarize information across the video-clip
  - Video-level prompts to model the data distribution
  - Textual prompts to enhance text description



Background

- Contrastive Language Image Pretraining (CLIP)
  - Pretrained on 400M image-text pair
  - Strong generalization and zeroshot capabilities



- Motivation for adapting CLIP to Videos
  - Lack of video-language data
  - Much larger computational requirements
  - Existing methods XCLIP, ActionCLIP

Challenges

- Trade-off in adopting Image Language models to Videos
  - Finetuning the backbone reduces zeroshot performance
  - Frozen backbone results in poor supervised performance

- State-of-the-art methods such as XCLIP tend to have separate training schemes for supervised and zeroshot settings
  - Essentially two different models!!!
  - Might as well have separate models for supervised/zeroshot

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Problem Formulation

Method	Epochs	Frames	K400 Supervised	HMDB51 Zeroshot	UCF101 Zeroshot	Trainable Parameters
XCLIP (Supervised)	30	8	82.3	41.4	67.9	131.5 M
XCLIP (Zeroshot)	10	32	78.2	44.6	72.0	131.5 M

- *Can we build a single model under a unified training scheme?*

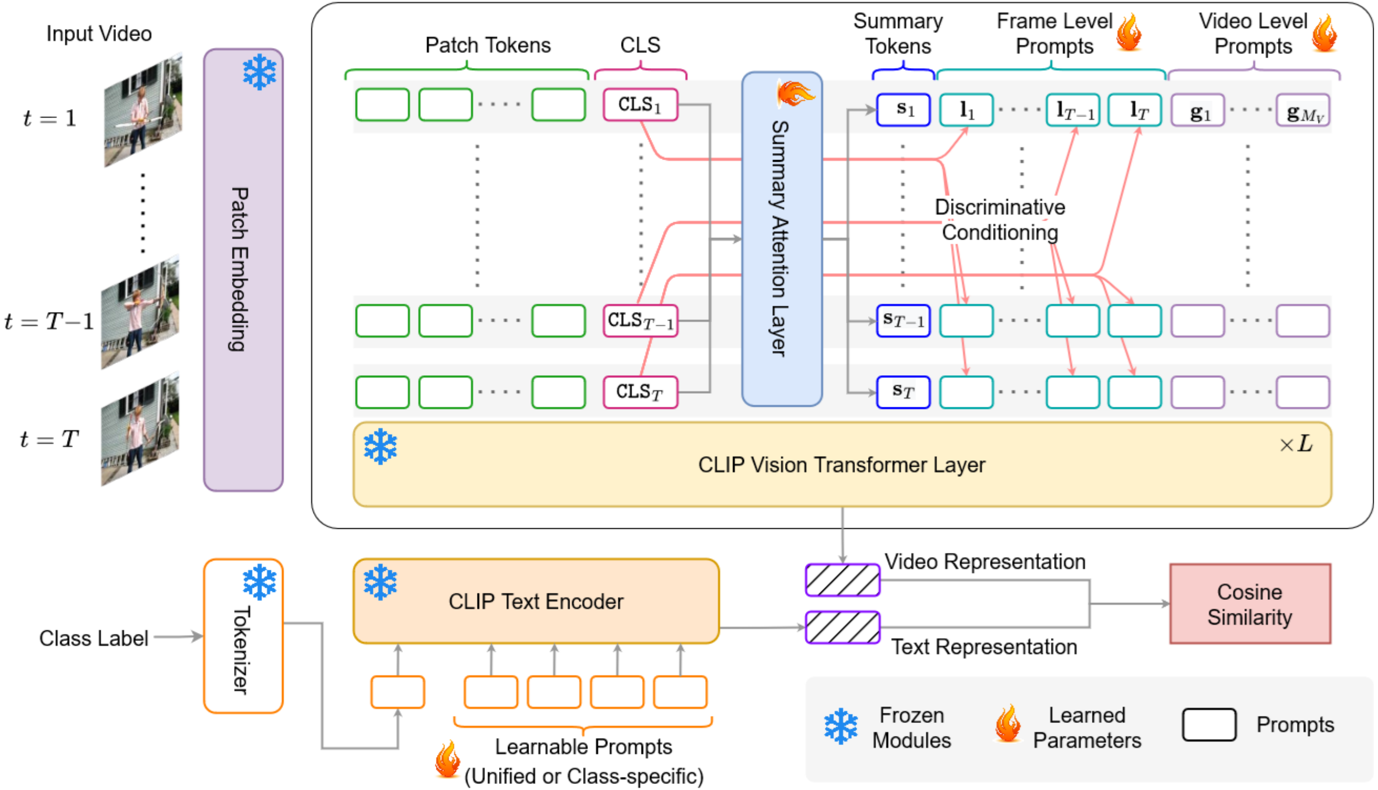
Methodology

- Freeze the backbone to retain CLIP generalization
- Introduce multimodal prompts to improve representation towards new dataset
- Condition prompts to model temporal information

Methodology

- Visual prompts at three levels of granularity
  - Local prompts to model frame-level information
  - Summary prompts to model a summarized representation across frames
  - Global prompts to model dataset distribution
- Textual prompts to enhance text representation

Methodology



Results

Method	Epochs	Frames	K400 Supervised	HMDB51 Zeroshot	UCF101 Zeroshot	Trainable Parameters
XCLIP (Supervised)	30	8	82.3	41.4	67.9	131.50 M
XCLIP (Zeroshot)	10	32	78.2	44.6	72.0	131.50 M
Vita-CLIP (Unified)	30	8	80.5	48.6	75.0	38.88 M

Conclusion

- We propose a *unified* model for both supervised and zeroshot settings

- We achieve state-of-the-art zeroshot performance, while still comparable in supervised setting

- We optimize a much smaller number of parameters

Thank You