



清华大学  
Tsinghua University



JD.COM

# DiscoVLA: Discrepancy Reduction in Vision, Language, and Alignment for Parameter-Efficient Video-Text Retrieval

Leqi Shen<sup>1,2,4\*</sup> Guoqiang Gong<sup>5†</sup> Tianxiang Hao<sup>1,2</sup> Tao He<sup>6,7</sup> Yifeng Zhang<sup>5</sup>  
Pengzhang Liu<sup>5</sup> Sicheng Zhao<sup>2‡</sup> Jungong Han<sup>3</sup> Guiguang Ding<sup>1,2‡</sup>

<sup>1</sup> School of Software <sup>2</sup> BNRist <sup>3</sup> Department of Automation, Tsinghua University

<sup>4</sup> Hangzhou Zhuoxi Institute of Brain and Intelligence

<sup>5</sup> JD.com <sup>6</sup> GRG Banking Equipment Co., Ltd. <sup>7</sup> South China University of Technology



# Introduction

Video-Text **R**etrieval.

- Matching video content with relevant textual descriptions or vice versa.
- Recent studies focus on extending **image-text** pretrained CLIP to VTR.

Limitations.

- Fully fine-tuning these pretrained models requires updating numerous parameters for every dataset, resulting in **large storage overhead**

In this work, we focus on parameter-efficient fine-tuning TVR.

- Targeting strong performance with minimal trainable parameters



# Introduction

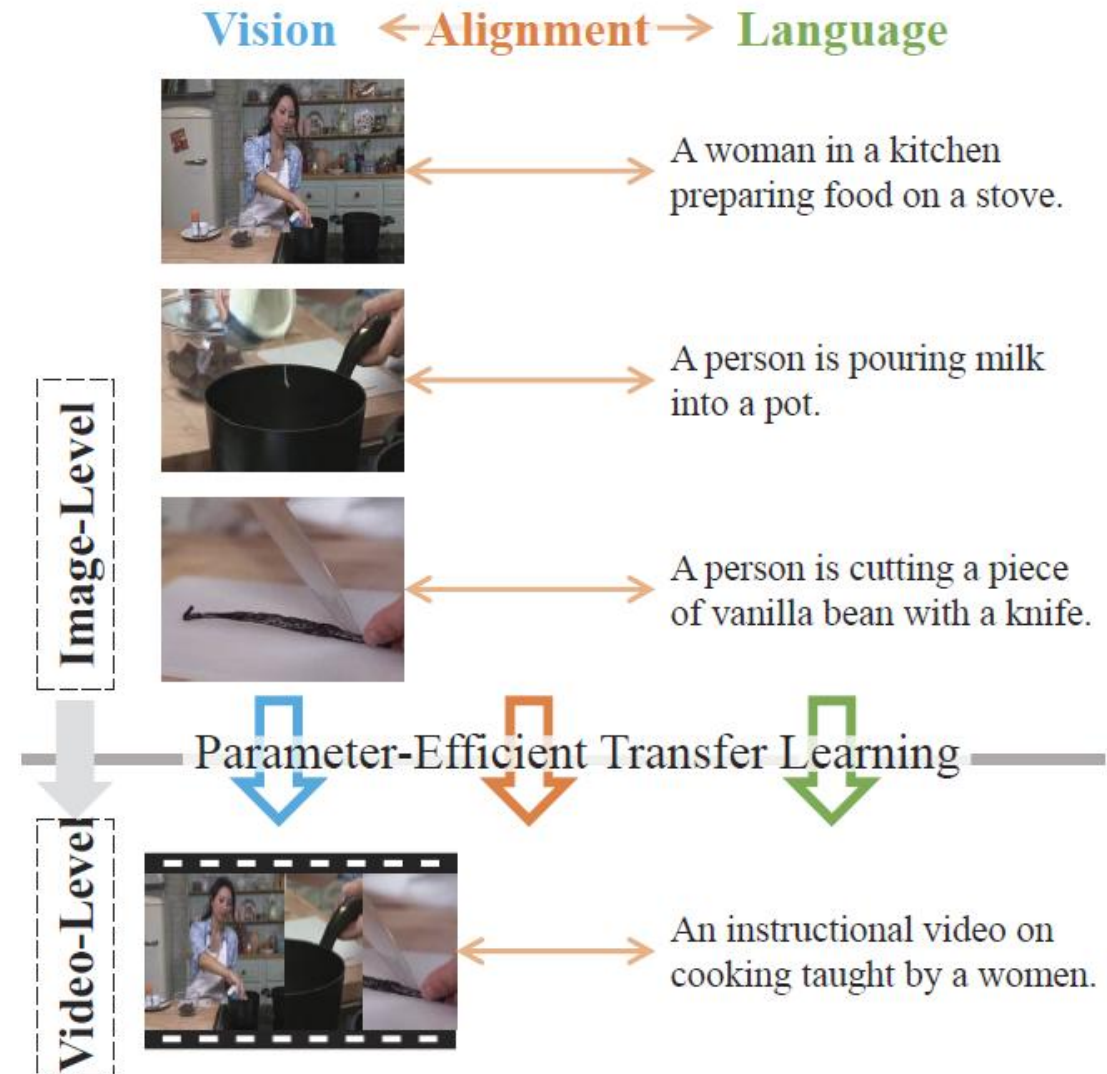
Challenges in transferring vision-language (VL) matching from image-text to video-text tasks.

Image-text pretraining.

- Focuses on matching images with image captions
- Image-level VL matching

Video-text retrieval.

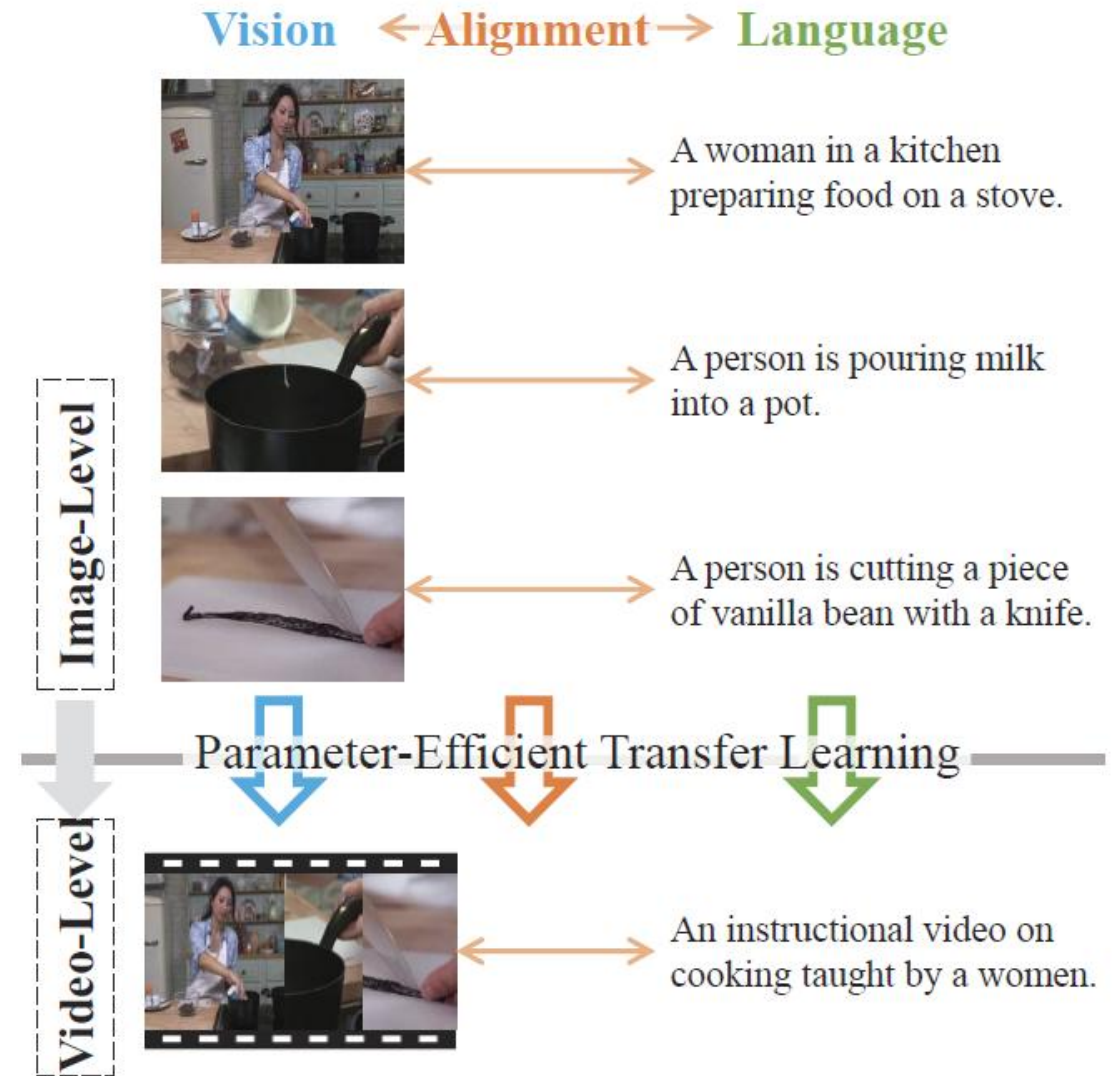
- Focuses on matching videos with video captions
- Video-level VL matching



# Introduction

The discrepancies between image-level and video-level VL matching across three key aspects:

- **Vision:** videos introduce a temporal dimension, which is absent in images.
- **Language:** the distinction arises from the varying levels of granularity
- **Alignment:** video-level alignment is inherently more complex due to intricate spatio-temporal relationships.



# Introduction

Current methods struggle to fully address the discrepancies.

To tackle these challenges, we introduce DiscoVLA  
aims to simultaneously reduce all significant discrepancies.

- We introduce IVFusion to fuse image- and video-level features for both vision and language gaps.
- PImgAlign is introduced for fine-grained image-level alignment.
- We introduce AlignDistill to minimize alignment gaps.

Method	Image-Level $\xrightarrow{\text{PETL}}$ Video-Level		
	Vision	Language	Alignment
Previous Methods	✓	✗	✗
Our DiscoVLA	✓	✓	✓





⊗ CLS Token of Vision Encoder   ⊛ EOS Token of Text Encoder   - - -> Utilized **ONLY** during the training phase.

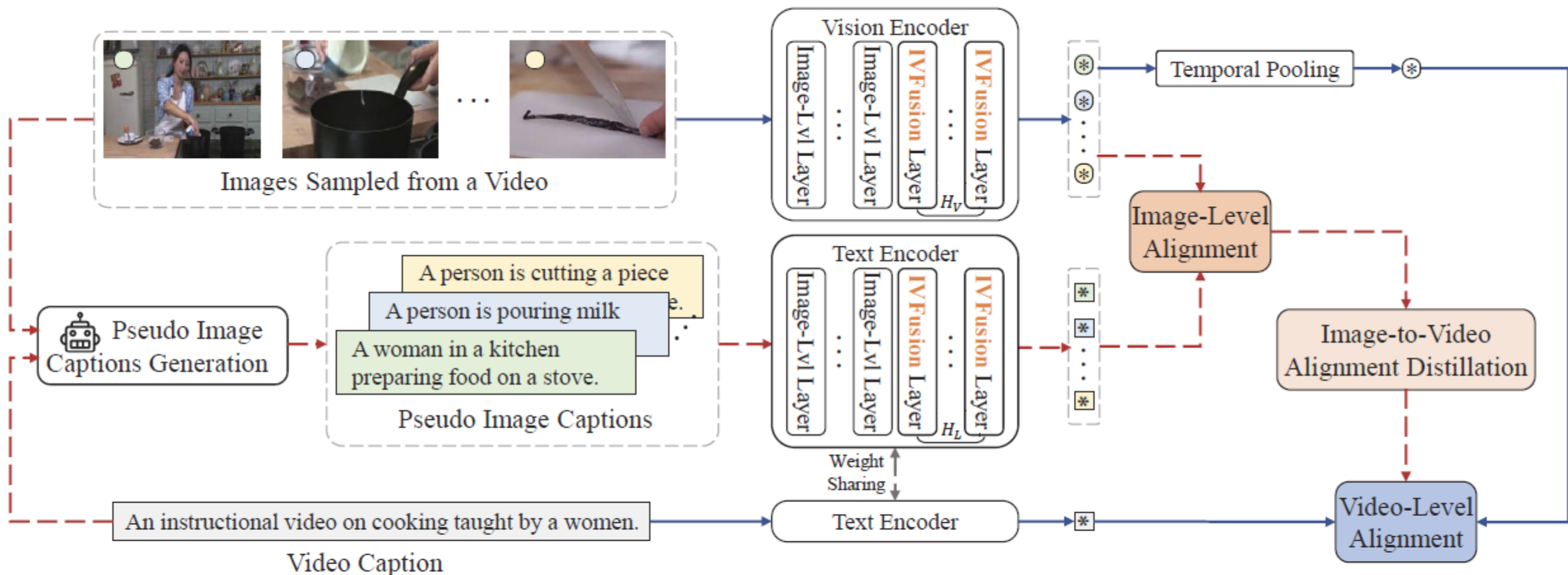


Figure 2. The overall framework of DiscoVLA. Initially, we generate pseudo image captions for each sampled image. In both vision and text encoders, we utilize image-level layers to acquire pretrained image-level knowledge and employ IVFusion layers to enhance spatio-temporal information. The single video caption is encoded through the text encoder, utilizing IVFusion layer as image-level (image-lvl) layer. Finally, AlignDistill is applied to distill image-level alignment to video-level alignment. For fair comparisons with previous methods, we do not generate pseudo image captions during the inference phase.

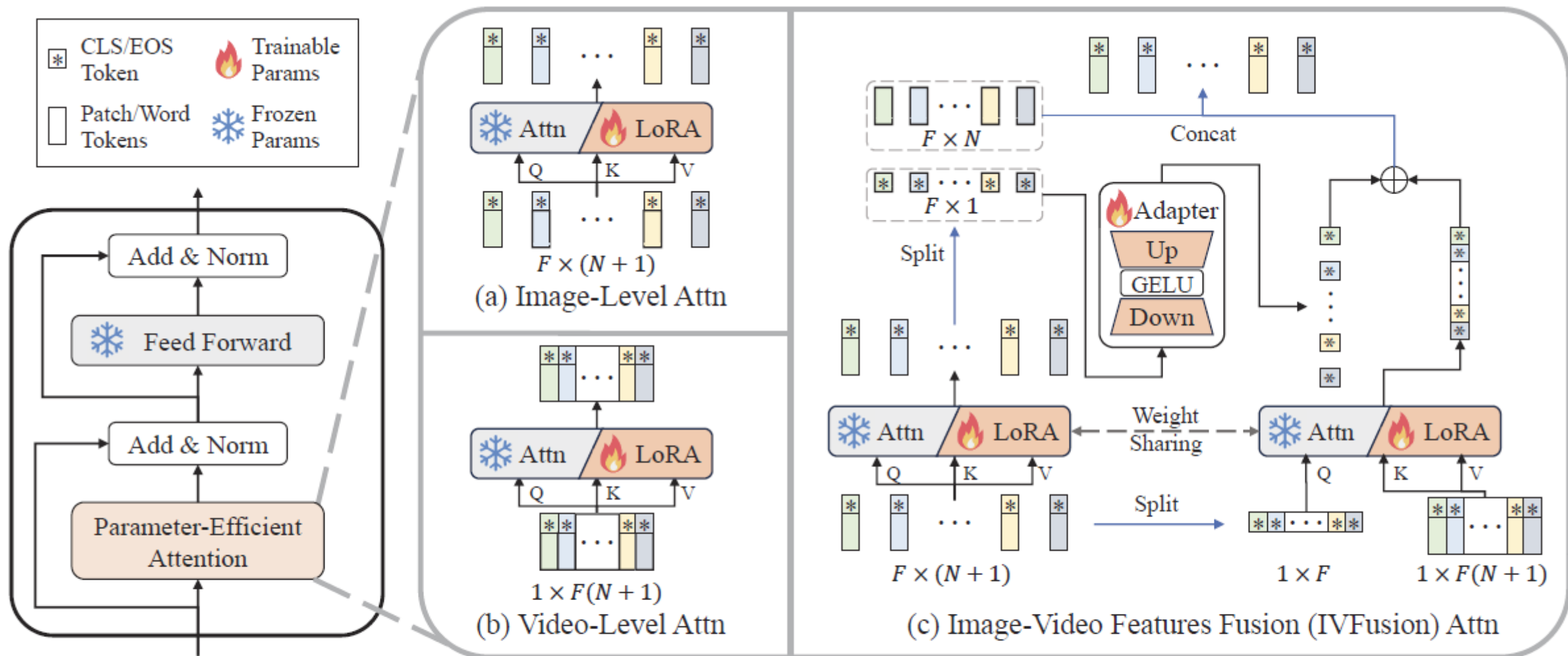


Figure 3. Illustration of the encoder layers for vision and text encoders. (a) Image-Level Attn operates on each of the  $F$  images or image captions individually. (b) Video-Level Attn concatenates tokens across all sampled images or image captions. (c) IVFusion Attn employs a lightweight adapter to integrate the efficiency of Image-Level Attn with the effectiveness of Video-Level Attn. Here, we illustrate the application of IVFusion within the vision encoder. The text encoder adopts the same approach for processing pseudo image captions.

# Methodology

## Pseudo Image-Level Alignment (PImgAlign)

- PImgAlign learns fine-grained image-level alignment via the decomposition of video.
- We utilize LLaVA-NeXT to generate pseudo image captions. For each sampled image from a video, we prompt the model using the relevant video caption for guidance:

The provided image is a frame sampled from the video, which describes *{video caption}*.  
Based on the video's content, provide a caption for the provided image.

- Given image features and pseudo image caption features from a paired video and video caption, we introduce a fine-grained image-level similarity:

$$\text{Sim}_{ij-t2v}^{\text{img}} = \frac{1}{F} \sum_{n=1}^F \max_{1 \leq m \leq F} t_{i|n}^{\text{img}T} v_{j|m}^{\text{img}},$$

$$\text{Sim}_{ij-v2t}^{\text{img}} = \frac{1}{F} \sum_{m=1}^F \max_{1 \leq n \leq F} t_{i|n}^{\text{img}T} v_{j|m}^{\text{img}},$$

$$\text{Sim}_{ij}^{\text{img}} = \frac{1}{2} [\text{Sim}_{ij-t2v}^{\text{img}} + \text{Sim}_{ij-v2t}^{\text{img}}],$$





# Methodology

## Image-to-Video Alignment Distillation (AlignDistill)

- AlignDistill distills image-level alignment knowledge into video-level alignment, thus mitigating alignment discrepancies.
- We optimize the Kullback-Leibler divergence between image-level similarity and video-level similarity:

$$\begin{aligned}\mathcal{S}_{t2v}^{\text{img}} &= [s_{i1}^{\text{img}}, \dots, s_{iN_v}^{\text{img}}], \mathcal{S}_{v2t}^{\text{img}} = [s_{1i}^{\text{img}}, \dots, s_{N_ti}^{\text{img}}], \\ \mathcal{S}_{t2v}^{\text{vid}} &= [s_{i1}^{\text{vid}}, \dots, s_{iN_v}^{\text{vid}}], \mathcal{S}_{v2t}^{\text{vid}} = [s_{1i}^{\text{vid}}, \dots, s_{N_ti}^{\text{vid}}], \\ \mathcal{L}_{KL} &= \frac{1}{2}(\text{KL}(\mathcal{S}_{t2v}^{\text{img}} \parallel \mathcal{S}_{t2v}^{\text{vid}}) + \text{KL}(\mathcal{S}_{v2t}^{\text{img}} \parallel \mathcal{S}_{v2t}^{\text{vid}})).\end{aligned}$$



Method	# Params (M)	Text-to-Video					Video-to-Text				
		R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓	R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓
CLIP (ViT-B/32)											
Full fine-tuning	123.54	43.1	70.4	80.8	194.3	16.2	43.1	70.5	81.2	194.8	12.4
Prompt [27]	0.08	40.4	66.3	77.3	184.0	16.7	42.2	69.7	79.2	191.1	12.4
Adapter [20]	0.26	41.9	69.9	78.7	190.2	14.9	43.6	69.9	80.1	193.6	11.5
LoRA [21]	0.49	43.7	68.9	80.4	193.0	16.0	43.0	70.2	82.2	195.4	12.0
RAP [5]	1.06	44.8	71.4	81.5	197.7	14.4	44.0	71.9	82.4	198.3	10.1
VoP <sup>F+P</sup> [22]	0.4	43.5	69.3	79.3	192.1	14.8	43.6	71.2	81.2	196.0	11.0
VoP F+C [22]	14.10	44.7	70.5	79.2	194.4	16.2	42.1	70.0	80.6	192.7	13.4
DGL [58]	0.83	44.6	69.9	80.3	194.8	16.3	44.5	70.7	80.6	195.8	11.5
<b>DiscoVLA</b>	0.56	<b>47.0</b>	<b>73.0</b>	<b>82.8</b>	<b>202.8</b>	<b>14.1</b>	<b>47.7</b>	<b>73.6</b>	<b>83.6</b>	<b>204.9</b>	<b>10.0</b>
CLIP (ViT-B/16)											
MV-Adapter [25]	3.6	46.0	72.0	82.1	200.1	-	45.6	74.0	83.8	203.4	-
RAP [5]	1.06	46.5	73.9	82.0	202.4	12.1	45.3	<b>76.4</b>	84.8	206.5	9.1
VoP <sup>F+P</sup> [22]	0.4	47.1	72.4	81.8	201.3	12.9	-	-	-	-	-
VoP <sup>F+C</sup> [22]	14.10	47.7	72.4	82.2	202.3	12.0	-	-	-	-	-
DGL [58]	0.83	48.3	71.8	80.6	200.7	13.4	45.7	74.0	82.9	202.6	10.9
TempMe [44]	0.50	49.0	74.4	83.3	206.7	<b>11.9</b>	47.6	75.3	<b>85.4</b>	208.3	9.0
<b>DiscoVLA</b>	0.56	<b>50.5</b>	<b>75.6</b>	<b>83.8</b>	<b>209.9</b>	12.1	<b>49.2</b>	76.0	84.7	<b>209.9</b>	<b>8.6</b>

- Our DiscoVLA with CLIP (ViT-B/32) as the backbone achieves **47.0%** R@1 in the text-to-video task (t2v) and **47.7%** R@1 in the video-to-text task (v2t), significantly outperforming previous methods.
- When using CLIP (ViT-B/16), DiscoVLA achieves improvements of **2.2%** R@1 and **7.5%** R@sum.
















A man prepares food for his dog to eat - the dog is in a dress.	It is an advertisement of car.	Girl and boy hang out on the beach.	A man is doing home improvement.
 <p>A dog wearing a purple dress is sitting on the floor.</p>	 <p>A group of people are walking around a car dealership, with a white Audi on display.</p>	 <p>A girl sitting on a beach chair while another person is buried in the sand.</p>	 <p>A man is wearing a white shirt with the words "Drain &amp; Clean" on it.</p>
 <p>A stainless steel oven with a glass door and four knobs on top.</p>	 <p>A man and woman are smiling and talking to another man in a business setting.</p>	 <p>A girl and a boy are sitting on the beach, enjoying the view of palm trees and a green flag.</p>	 <p>A man is working on the ceiling of a room.</p>
 <p>A bowl of food with sliced sweet potatoes, spinach, and other ingredients.</p>	 <p>A group of people in suits standing in a car showroom.</p>	 <p>Two people sitting on a lifeguard chair with their arms raised in the air.</p>	 <p>A man is working on the ceiling of a room.</p>
 <p>A dog wearing a dress stands in a kitchen.</p>	 <p>A sleek black Audi car is parked in a showroom.</p>	 <p>A couple sitting on a lifeguard chair on the beach.</p>	 <p>A man is installing a pipe in the corner of a room.</p>
(a)	(b)	(c)	(d)

Figure 4. Visualization of discrepancies between video-level and image-level data. Each example consists of a paired video and video caption, with four sampled images from the video. Video captions are highlighted with orange solid lines, and pseudo image captions generated by LLaVA-NeXT [35] are indicated with green dashed lines. While video captions convey the general context of the video, image captions focus on the detailed context of each individual frame.



THANKS!

