# VINDLU: A Recipe for Effective Video-and-Language Pretraining

Feng Cheng[1], Xizi Wang[2], Jie Lei[1], David Crandall[2], Mohit Bansal[1], Gedas Bertasius[1]
[1]University of North Carolina at Chapel Hill, [2]Indiana University

CVPR 2023

**Poster Session: WED-AM-239**

# Motivation

## Video-language pretraining

- Expensive to train

| Model | V100-GPU days |
|---|---|
| ALL-in-one | 448 |
| LAVENDER | 640 |
| CLIP-ViP | 984 |

- Complex architectures

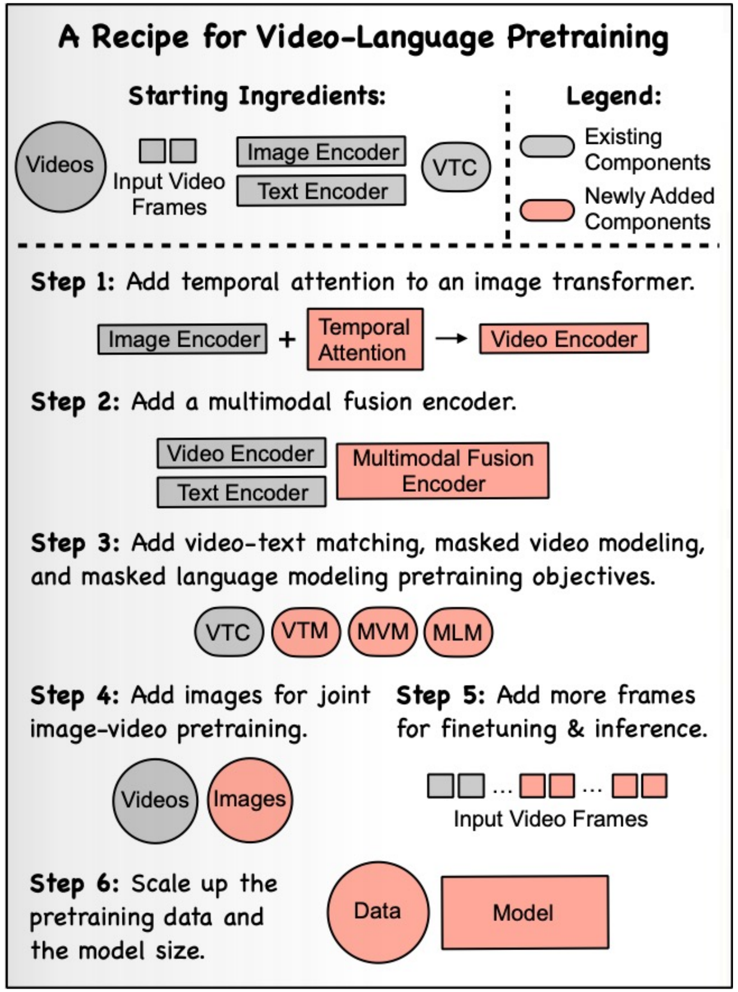
| Method | Model Design | | | Pretraining Data | | | #Frames | | |
|---|---|---|---|---|---|---|---|---|---|
| | Temporal Modeling | Multimodal Fusion | Pretraining Objectives | Dataset | Size | Modality | PT | FT | Eval |
| UniVL [48] | Joint Att. [5] | 2-layer TR | VTC+VTM+MLM+MFM+LM | HT | 136M | V | 48 | 48 | 48 |
| VideoCLIP [75] | 1D-Conv+TR | ✗ | VTC | HT | 136M | V | 32 | 32 | 32 |
| ClipBert [32] | Mean Pooling | BERT | MLM+VTM | COCO+VG | 0.2M | I | 1 | 16 | 16 |
| Frozen [2] | Temp. Attn [5] | ✗ | ITC | C5M | 5M | I+V | $1 \rightarrow 4$ | 4 | 4 |
| MERLOT [86] | Joint Attn | RoBERTa | VTC+MLM+FOM | YT | 180M | V | 16 | 16 | 16 |
| VIOLET [19] | Window Attn [44] | BERT | VTC+VTM+MLM+MVM | YT+C5M | 185M | I+V | 4 | 5 | 5 |
| MV-GPT [59] | Joint Attn | 2-layer TR | MLM+LM | HT | 136M | V | - | - | - |
| ALL-in-one [67] | Token Rolling [67] | ViT | VTC+VTM+MLM | HT+W2 | 172M | V | 3 | 3 | 9 |
| Singularity [31] | Late Temp. Attn | 3-layer TR | VTC+VTM+MLM | C17M | 17M | I+V | $1 \rightarrow 4$ | 4 | 12 |
| LAVENDER [38] | Window Attn [44] | BERT | MLM | C17M+IN | 30M | I+V | 4 | 5 | 5 |
| OmniVL [69] | Temp. Attn | 2×BERT | VTC+VTM+LM | C17M | 17M | I+V | $1 \rightarrow 8$ | 8 | 8 |
| ATP [6] | ✗ | ✗ | VTC | CLIP | 400M | I | 1 | 16 | 16 |
| CLIP4Clip [49] | Late TR | ✗ | VTC | CLIP | 400M | I | 1 | 12 | 12 |
| ECLIPSE [40] | Late TR | ✗ | VTC | CLIP | 400M | I+A | 1 | 32 | 32 |
| CLIP2TV [21] | CLIP | 4-layer TR | VTC+VTM | CLIP | 400M | I | 1 | 12 | 12 |
| CLIP-Hitchhiker [3] | Late Attn | ✗ | VTC | CLIP | 400M | I | 1 | 16 | 120 |
| CLIP-ViP [77] | Prompt Attn [77] | ✗ | VTC | CLIP | 500M | I+V | $1 \rightarrow 12$ | 12 | 12 |

**TR**: Transformer; **Late**: Late fusion; **Attn**: Attention. **V**: Video; **I**: Image; **A**: Audio; $1 \rightarrow 4$: 1 frame for stage-1 training and 4 frames for stage-2. **VTC**: Video-text contrastive; **VTM**: Video-text matching; **MLM**: Masked language modeling; **MFM**: Masked frame modeling; **LM**: Language modeling. **HT**: HowTo100M [51]; **C5M, C17M**: see supplementary; **YT**: YT-Temporal [86]; **W2**: WebVid-2M [2]; **COCO**: [39], **VG**: Visual Genome [30]; **IN**: An internal dataset.

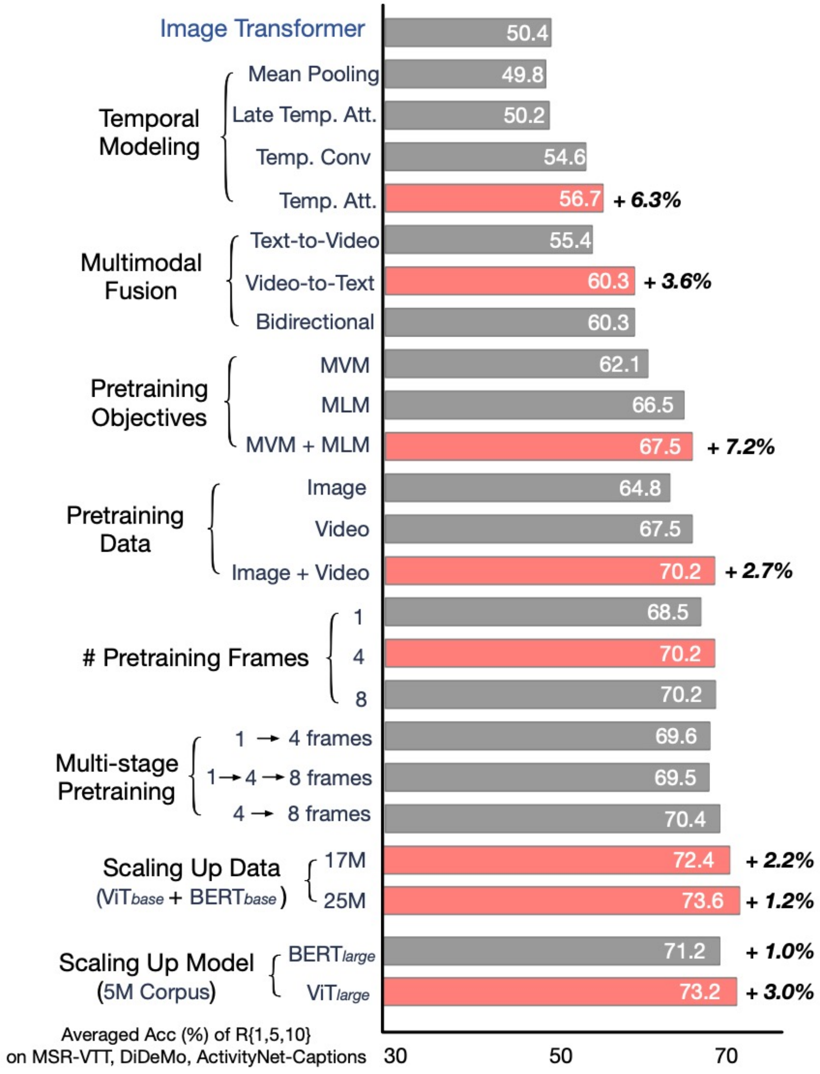
➢ **Goal: Make VidL pretraining more efficient, effective and accessible.**

# A Recipe for Effective VidL Pretraining

- We start with image and text encoders trained on video-text pairs using a contrastive loss.
- We then progressively add more components while studying the importance of each component.
- Using our empirical insights, we then develop a step-by-step recipe for effective VidL pretraining.



# Our Recipe: VindLU




Our final recipe outperforms the original baseline by 23.2%

# Experiments

Check our code!

VindLU achieves state-of-the-art results on 9 video-language benchmarks.

## Text-to-Video Retrieval

| Method | Pretrain | | | MSRVTT | | | | DiDeMo | | | | ActivityNet-Captions | | | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #Data | #Frames | Time | R1 | R5 | R10 | Avg | R1 | R5 | R10 | Avg | R1 | R5 | R10 | Avg | |
| ClipBERT [32] | 5.4M | 1 | 32 | 22.0 | 46.8 | 59.9 | 42.9 | 20.4 | 48.0 | 60.8 | 43.1 | 21.3 | 49.0 | 63.5 | 44.6 | 43.5 |
| VideoCLIP [75] | 136M | 960 | 8 | 30.9 | 55.4 | 66.8 | 51.0 | - | - | - | - | - | - | - | - | - |
| Frozen [2] | 5M | $1 \rightarrow 4$ | 35* | 31.0 | 59.5 | 70.5 | 53.7 | 34.6 | 65.0 | 74.7 | 58.1 | - | - | - | - | - |
| ALPRO [34] | 5M | 8 | 24* | 33.9 | 60.7 | 73.2 | 55.9 | 35.9 | 67.5 | 78.8 | 60.7 | - | - | - | - | - |
| VIOLET [19] | 138M | 4 | 83 | 34.5 | 63.0 | 73.4 | 57.0 | 32.6 | 62.8 | 74.7 | 56.7 | - | - | - | - | - |
| All-in-one [67] | 138M | 3 | 448 | 37.9 | 68.1 | 77.1 | 61.0 | 32.7 | 61.4 | 73.5 | 55.9 | 22.4 | 53.7 | 67.7 | 47.9 | 54.9 |
| LAVENDER [38] | 30M | 4 | 640 | 40.7 | 66.9 | 77.6 | 61.7 | 53.4 | 78.6 | 85.3 | 72.4 | - | - | - | - | - |
| Singularity [31] | 17M | $1 \rightarrow 4$ | 29 | 42.7 | 69.5 | 78.1 | 63.4 | 53.1 | 79.9 | 88.1 | 73.7 | 48.9 | 77.0 | 86.3 | 70.7 | 69.3 |
| OmniVL [69] | 17M | $1 \rightarrow 8$ | 169* | **47.8** | **74.2** | **83.8** | **68.6** | 52.4 | 79.5 | 85.4 | 72.4 | - | - | - | - | - |
| CLIP4Clip [49] | 400M | 1 | 768* | 44.5 | 71.4 | 81.6 | 65.8 | 42.8 | 68.5 | 79.2 | 63.5 | 40.5 | 72.4 | 83.4 | 65.4 | 64.9 |
| ECLIPSE [40] | 400M | 1 | 768* | - | - | - | – | 44.2 | - | - | - | 45.3 | 75.7 | 86.2 | 69.1 | - |
| CLIP-Hhiker [3] | 400M | 1 | 768* | 47.7 | 74.1 | 82.9 | 68.6 | - | - | - | - | 44.0 | 74.9 | 86.1 | 68.3 | - |
| CLIP-ViP [77] | 500M | $1 \rightarrow 12$ | 984* | 54.2 | 77.2 | 84.8 | 72.1 | 50.5 | 78.4 | 87.1 | 72.0 | 53.4 | 81.4 | 90.0 | 74.9 | 73.0 |
| VINDLU | 5M | 4 | 15 | 43.8 | 70.3 | 79.5 | 64.5 | 54.6 | 81.3 | 89.0 | 75.0 | 51.1 | 79.2 | 88.4 | 72.9 | 70.8 |
| | 17M | | 38 | 45.3 | 69.9 | 79.6 | 64.9 | 59.2 | 84.1 | 89.5 | 77.6 | 54.4 | 80.7 | 89.0 | 74.7 | 72.4 |
| | 25M | | 82 | 46.5 | 71.5 | 80.4 | 66.1 | **61.2** | **85.8** | **91.0** | **79.3** | **55.0** | **81.4** | **89.7** | **75.4** | **73.6** |

| Method | #PT | SSv2-label | | SSv2-template | | Avg |
|---|---|---|---|---|---|---|
| | | R1 | R5 | R1 | R5 | |
| CLIP4Clip [49] | 400M | 43.1 | 71.4 | 77.0 | 96.6 | 77.9 |
| Singularity [31] | 17M | 47.4 | 75.9 | 77.6 | 96.0 | 80.0 |
| VINDLU | 5M | 51.2 | 78.8 | 82.2 | 98.9 | 82.7 |
| | 17M | 53.0 | 80.8 | **86.2** | 99.4 | **84.6** |
| | 25M | **53.1** | **81.8** | 83.3 | **100** | 84.4 |

## Video Question Answering

| Method | #PT | ANet | MSR-QA | MSR-MC | TVQA |
|---|---|---|---|---|---|
| ClipBERT [32] | 0.2M | - | 37.4 | 88.2 | - |
| ALPRO [34] | 5M | - | 42.1 | - | - |
| JustAsk [78] | 69M | 38.9 | 41.5 | - | - |
| VideoCLIP [75] | 136M | - | - | 92.1 | - |
| All-in-one [67] | 138M | - | 44.3 | 92.0 | - |
| MERLOT [86] | 180M | 41.4 | 43.1 | 90.9 | 78.7 |
| VIOLET [19] | 138M | - | 43.9 | 91.9 | - |
| Singularity [31] | 17M | 44.1 | 43.9 | 93.7 | - |
| OmniVL [69] | 17M | - | 44.1 | - | - |
| HERO [37] | 7.5M | - | - | - | 74.2 |
| FrozenBiLM [79] | 400M | 43.2 | 47.0 | - | 82.0 |
| VINDLU | 5M | 44.2 | 43.6 | 95.2 | **79.0** |
| | 17M | 44.6 | 43.8 | 96.7 | 78.8 |
| | 25M | **44.7** | **44.6** | **97.1** | **79.0** |

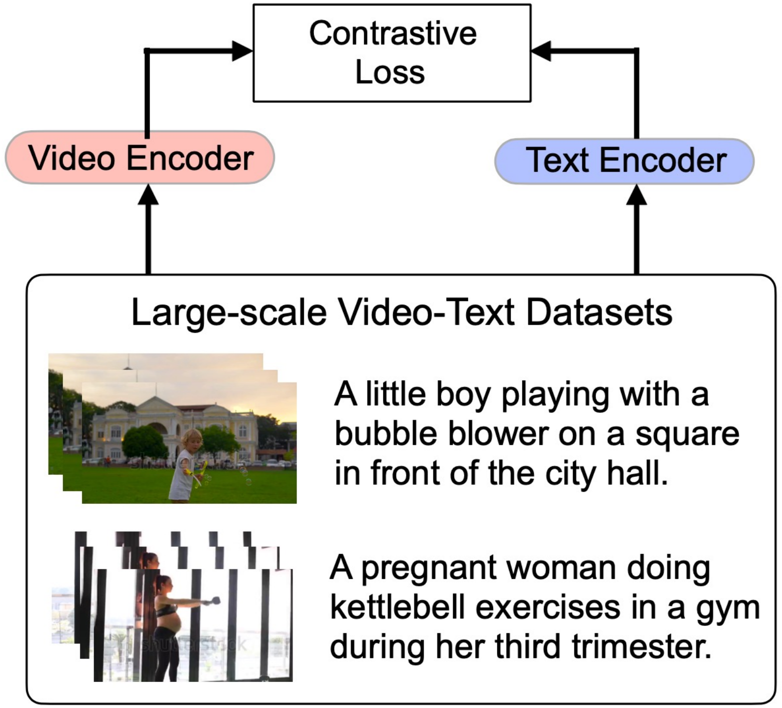
## Action Recognition

| Method | TimeSformer [5] | OmniVL [69] | VINDLU |
|---|---|---|---|
| Top-1 acc. | 78.0 | 79.1 | **80.1** |

# Details

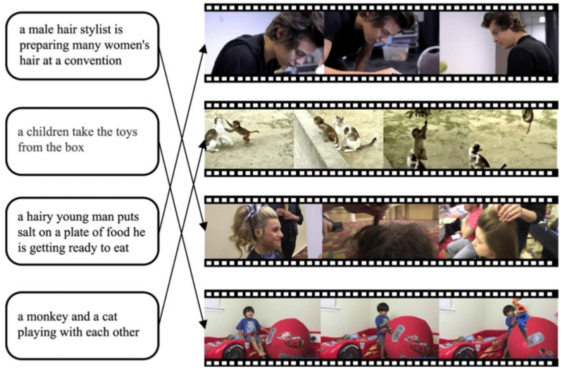
# Problem Statement

## Video-and-Language (VidL) Pretraining

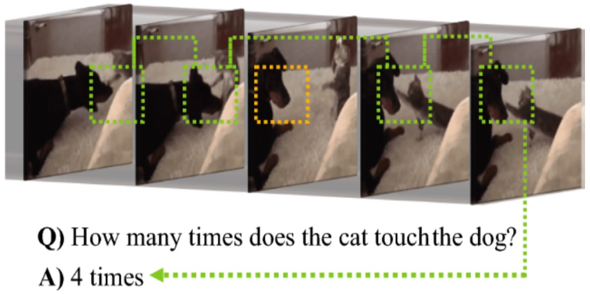



Finetuning

Text-to-video Retrieval




Video Question Answering




This scheme has been shown very effective for downstream VidL tasks.

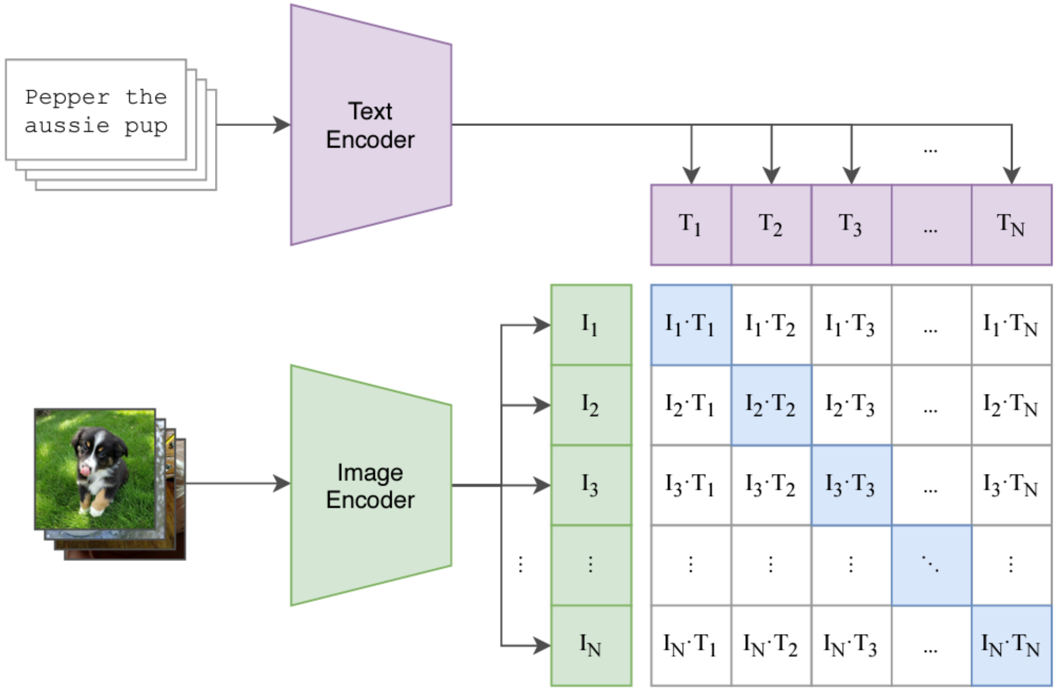
# Our Recipe: VindLU 🍲

## Step 0: Starting Ingredients

The image transformer baseline achieves 50.2% accuracy.

**Model:**
- Image Encoder: ViT-B/16.
- Text Encoder: BERT.

**Datasets:**
- Pretraining: WebVid-2M.
- Objective: VTC contrastive loss.
- Evaluation: Text-to-video retrieval on MSR-VTT, DiDeMo, ActivityNet.



# Our Recipe: VindLU 🍲

## Step 1: Temporal Modeling

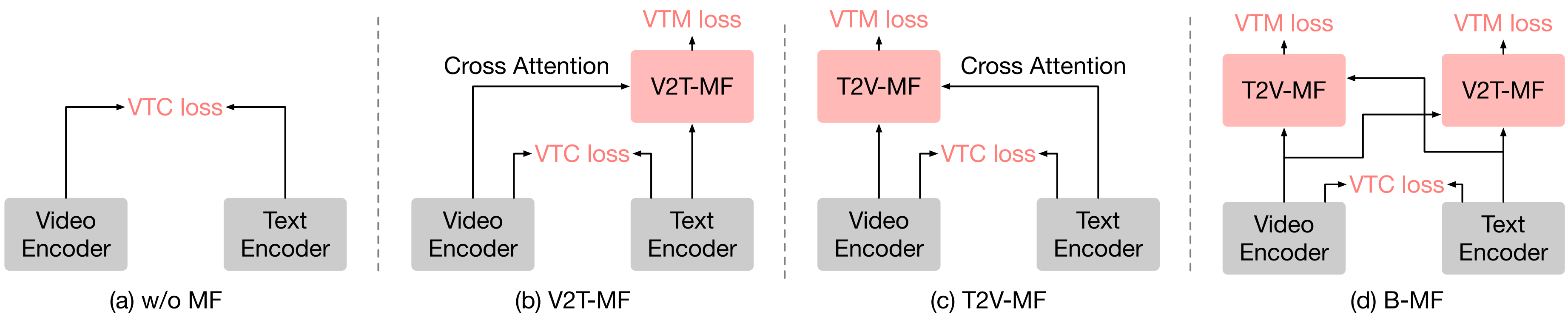
- **Mean Pooling:** the model averages independently computed frame-level scores.
- **L-TA:** adding 2 Transformer layers to an image encoder for temporal aggregation.
- **TC:** using 3D temporal convolutions for temporal modeling.
- **TA:** The divided space-time attention from TimeSformer inserted before spatial attention.

| | Mean Pooling | L-TA | TC | TA |
|---|---|---|---|---|
| acc.(%) | 49.8 | 50.2 | 54.6 | **56.7** |

# Our Recipe: VindLU

## Step 2: Multimodal Fusion Encoder

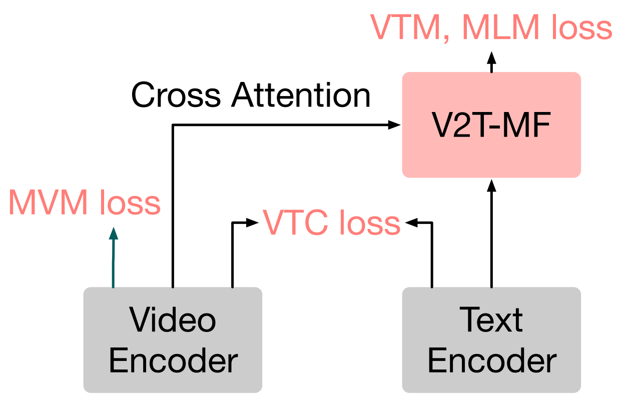
The purpose of the multimodal fusion encoder is to fuse multimodal cues from video and language.




| | w/o. MF | T2V-MF | V2T-MF | B-MF |
|---|---|---|---|---|
| acc.(%) | 56.7 | 55.4 | **60.3** | **60.3** |

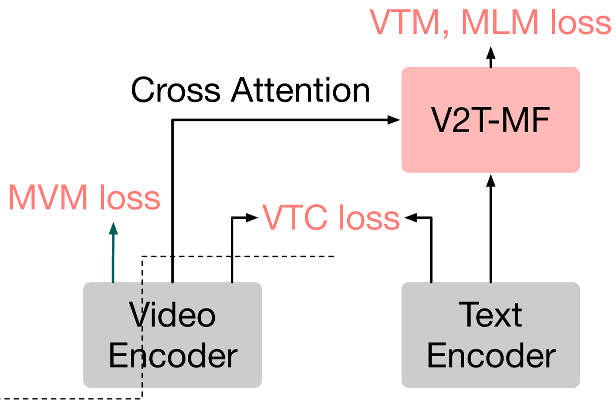
# Our Recipe: VindLU 🍲

## Step 3: Pretraining Objectives

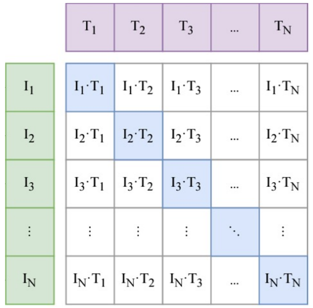


# Our Recipe: VindLU

## Step 3: Pretraining Objectives



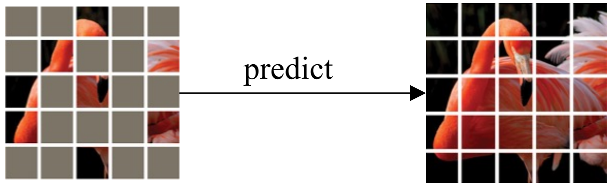

- **Video-to-Text Contrastive (VTC)**

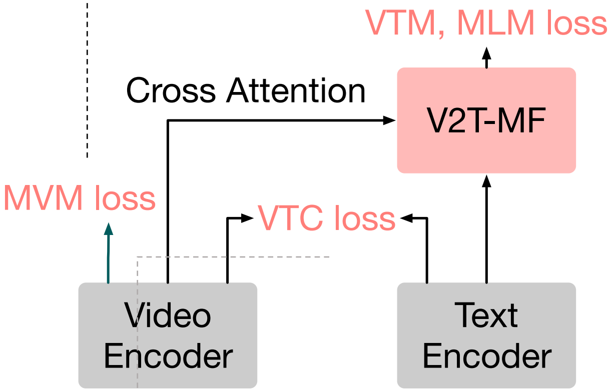


# Our Recipe: VindLU

## Step 3: Pretraining Objectives

- **Masked Video Modeling (MVM)**



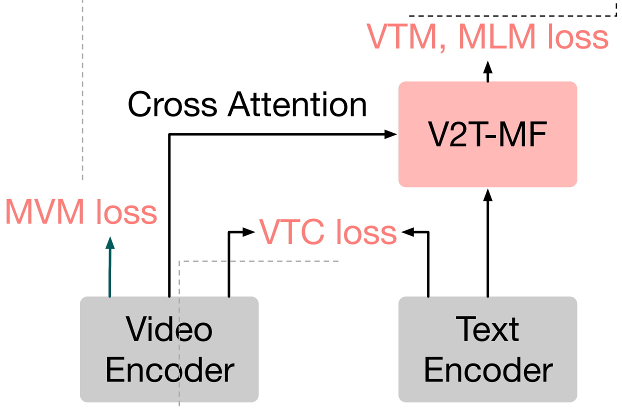

- **Video-to-Text Contrastive (VTC)**

| | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

# Our Recipe: VindLU 🍲

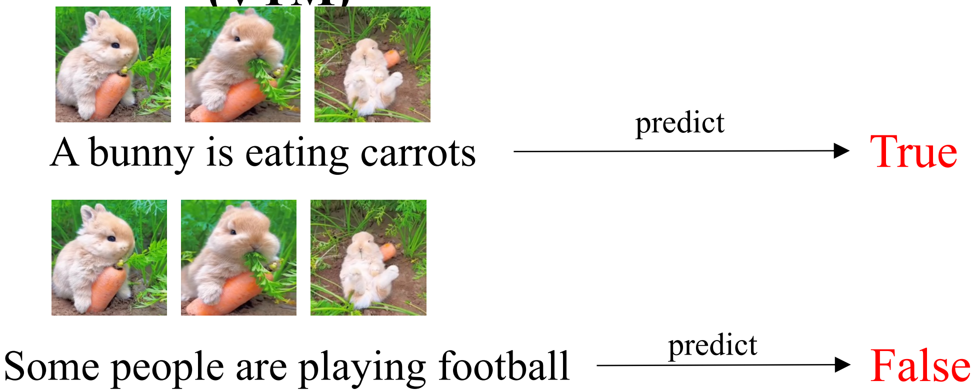
## Step 3: Pretraining Objectives

- Masked Video Modeling (MVM)

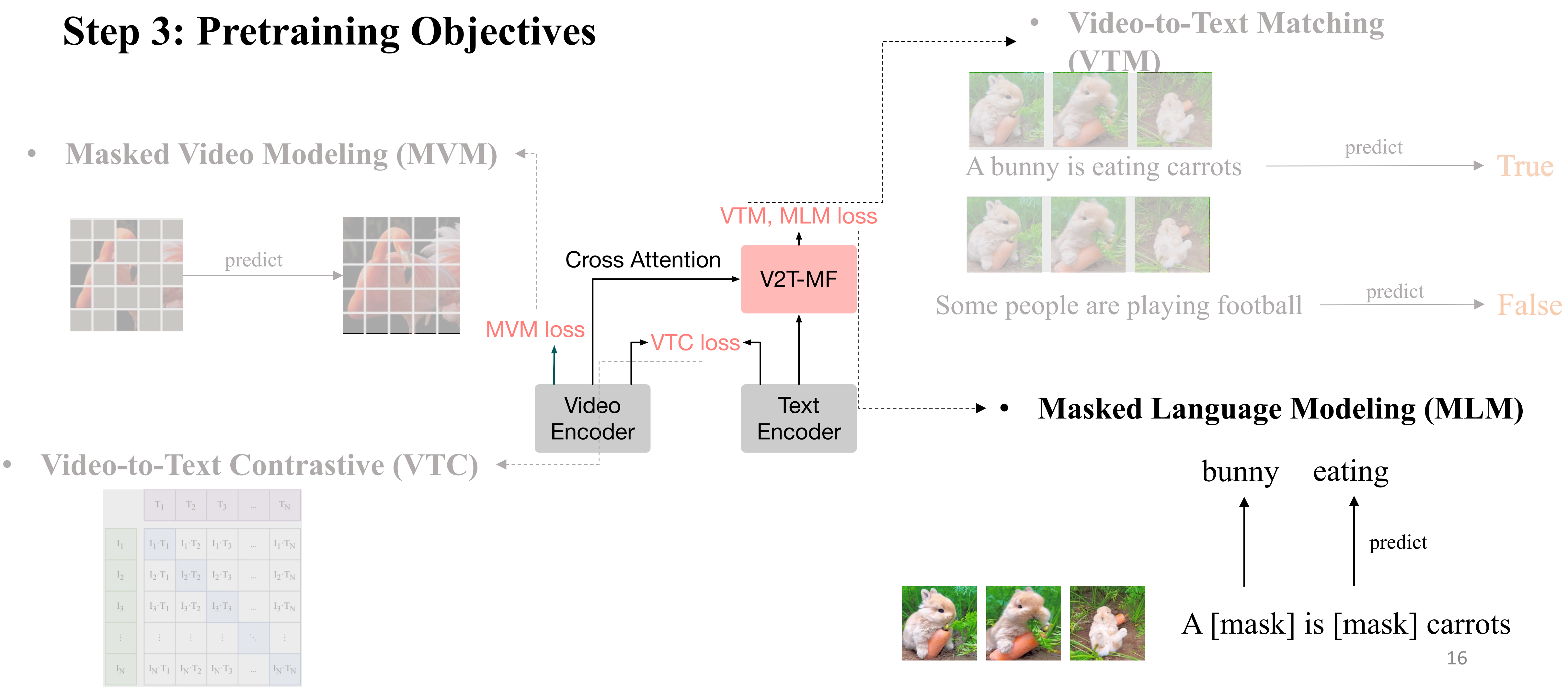

- Video-to-Text Contrastive (VTC)

- Video-to-Text Matching (VTM)



# Our Recipe: VindLU

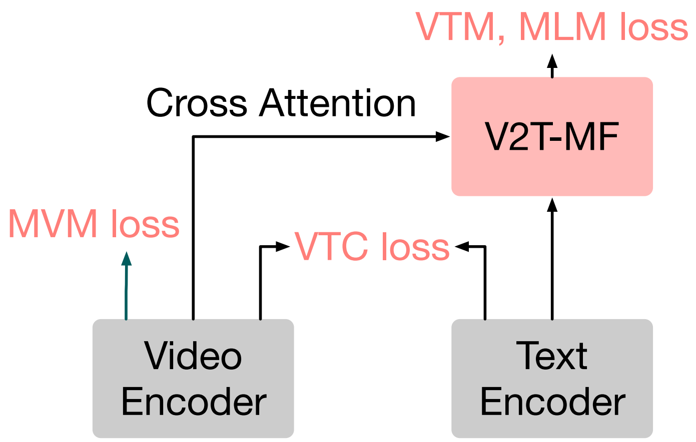
## Step 3: Pretraining Objectives



# Our Recipe: VindLU

## Step 3: Pretraining Objectives

- **VTC:** contrastive video-text loss objective.
- **VTM:** non-contrastive video-text matching classification objective attached to multimodal encoder.
- **MLM:** masked word token prediction loss.
- **MVM:** masked video token prediction loss.




| objectives | acc.(%) |
|---|---|
| VTC (Step 1) | 56.7 |
| VTC+VTM (Step 2) | 60.3 |
| VTC+VTM+MLM | 66.5 |
| VTC+VTM+MLM+MVM | **67.5** |

# Our Recipe: VindLU 🍲

## Step 4: Pretraining Data

| | Images | Videos | Images+Videos |
|---|---|---|---|
| acc.(%) | 64.8 | 67.5 | **70.2** |

Training Jointly on images and videos is beneficial.

| | 1 frame | 4 frames | 8 frames | 16 frames |
|---|---|---|---|---|
| acc.(%) | 68.5 | **70.2** | **70.2** | **70.2** |
| speedup | **4.6**× | 2.5× | 1.7× | 1× |

Pretraining on 4 frames is sufficient and provides large reduction in the computational cost.

| frames | 4 | $1 \rightarrow 4$ | $1 \rightarrow 4 \rightarrow 8$ | $4 \rightarrow 8$ |
|---|---|---|---|---|
| acc.(%) | 70.2 | 69.6 | 69.5 | **70.4** |
| speedup | **1.7**× | **1.7**× | 1.2× | 1× |

Multi-stage pretraining is not necessary.

## Our Recipe: VindLU 🍲

### Step 5: Finetuning and inference

- Finetuning

| # frames | 1 | 4 | 8 | 12 | 24 | 32 |
|---|---|---|---|---|---|---|
| acc.(%) | 65.5 | 68.1 | 69.2 | 70.2 | 70.1 | **70.5** |
| speedup | **22.4**× | 7.1× | 3.9× | 2.6× | 1.5× | 1.0× |

Finetuning on 12 frames provides a good tradeoff between accuracy and cost.

- Inference

| # frames | 12 | 24 | 32 | 64 |
|---|---|---|---|---|
| D/A acc.(%) | 73.4/70.4 | 73.0/72.1 | 72.7/72.6 | **73.8/72.8** |
| speedup | **10.6**× | 3.1× | 2.1× | 1× |

Inference with more frames yields slightly better results at larger computational cost.

# Our Recipe: VindLU 🍲

## Step 6: Scaling Up

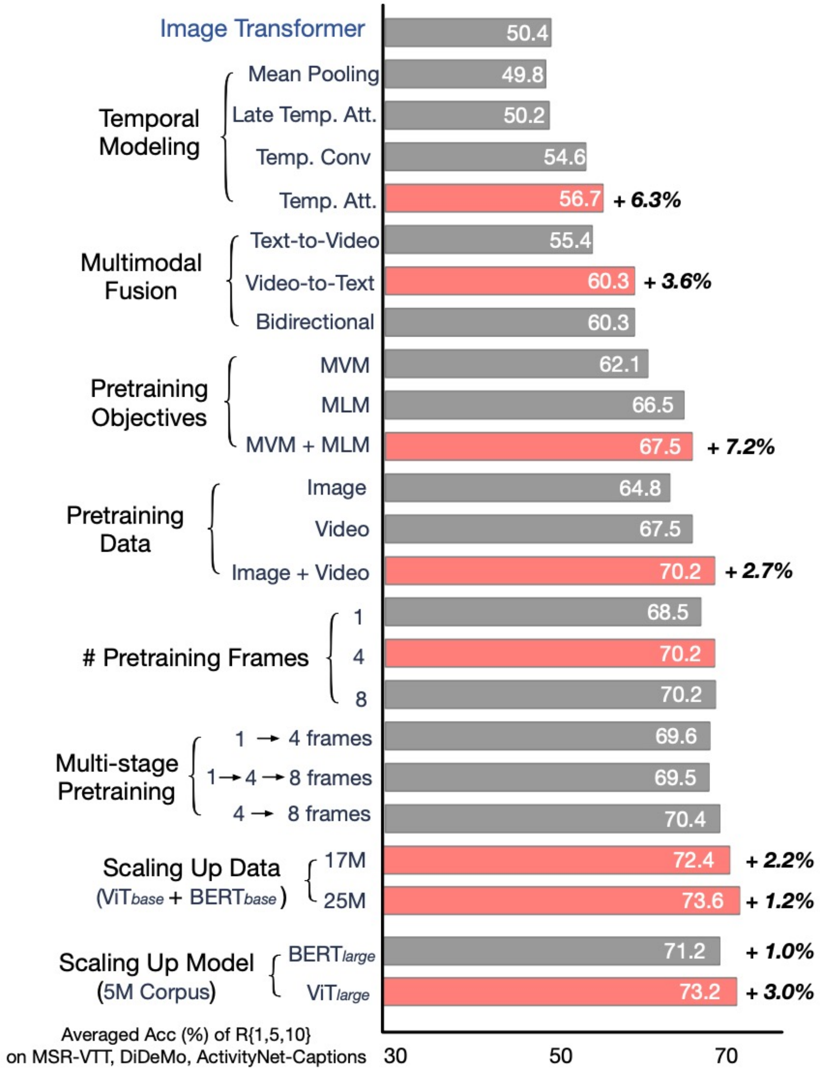
| # corpus | 5M | 17M | 25M |
|---|---|---|---|
| acc.(%) | 70.2 | 72.4 | **73.6** |

Scaling the corpus leads to 3.4% boost.

| encoders | base | $\text{ViT}_{large}$ | $\text{BERT}_{large}$ |
|---|---|---|---|
| acc.(%) | 70.2 | **73.2** | 71.2 |

Scaling the vision and text encoders lead to 3.0% and 1.0% boost respectively.

# Our Recipe: VindLU




Our final recipe outperforms the original baseline by 23.2%

# Experiments

## Text-to-Video Retrieval

| Method | Pretrain | | | MSRVTT | | | | DiDeMo | | | | ActivityNet-Captions | | | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #Data | #Frames | Time | R1 | R5 | R10 | Avg | R1 | R5 | R10 | Avg | R1 | R5 | R10 | Avg | |
| ClipBERT [32] | 5.4M | 1 | 32 | 22.0 | 46.8 | 59.9 | 42.9 | 20.4 | 48.0 | 60.8 | 43.1 | 21.3 | 49.0 | 63.5 | 44.6 | 43.5 |
| VideoCLIP [75] | 136M | 960 | 8 | 30.9 | 55.4 | 66.8 | 51.0 | - | - | - | - | - | - | - | - | - |
| Frozen [2] | 5M | $1 \rightarrow 4$ | 35* | 31.0 | 59.5 | 70.5 | 53.7 | 34.6 | 65.0 | 74.7 | 58.1 | - | - | - | - | - |
| ALPRO [34] | 5M | 8 | 24* | 33.9 | 60.7 | 73.2 | 55.9 | 35.9 | 67.5 | 78.8 | 60.7 | - | - | - | - | - |
| VIOLET [19] | 138M | 4 | 83 | 34.5 | 63.0 | 73.4 | 57.0 | 32.6 | 62.8 | 74.7 | 56.7 | - | - | - | - | - |
| All-in-one [67] | 138M | 3 | 448 | 37.9 | 68.1 | 77.1 | 61.0 | 32.7 | 61.4 | 73.5 | 55.9 | 22.4 | 53.7 | 67.7 | 47.9 | 54.9 |
| LAVENDER [38] | 30M | 4 | 640 | 40.7 | 66.9 | 77.6 | 61.7 | <u>53.4</u> | 78.6 | 85.3 | 72.4 | - | - | - | - | - |
| Singularity [31] | 17M | $1 \rightarrow 4$ | 29 | 42.7 | 69.5 | 78.1 | 63.4 | 53.1 | <u>79.9</u> | <u>88.1</u> | <u>73.7</u> | <u>48.9</u> | <u>77.0</u> | <u>86.3</u> | <u>70.7</u> | 69.3 |
| OmniVL [69] | 17M | $1 \rightarrow 8$ | 169* | **47.8** | **74.2** | **83.8** | **68.6** | 52.4 | 79.5 | 85.4 | 72.4 | - | - | - | - | - |
| CLIP4Clip [49] | 400M | 1 | 768* | 44.5 | 71.4 | 81.6 | 65.8 | 42.8 | 68.5 | 79.2 | 63.5 | 40.5 | 72.4 | 83.4 | 65.4 | 64.9 |
| ECLIPSE [40] | 400M | 1 | 768* | - | - | - | – | 44.2 | - | - | - | 45.3 | 75.7 | 86.2 | 69.1 | - |
| CLIP-Hhiker [3] | 400M | 1 | 768* | 47.7 | 74.1 | 82.9 | 68.6 | - | - | - | - | 44.0 | 74.9 | 86.1 | 68.3 | - |
| CLIP-ViP [77] | 500M | $1 \rightarrow 12$ | 984* | 54.2 | 77.2 | 84.8 | 72.1 | 50.5 | 78.4 | 87.1 | 72.0 | 53.4 | 81.4 | 90.0 | 74.9 | 73.0 |
| VindLU | 5M | 4 | 15 | 43.8 | 70.3 | 79.5 | 64.5 | 54.6 | 81.3 | 89.0 | 75.0 | 51.1 | 79.2 | 88.4 | 72.9 | 70.8 |
| | 17M | | 38 | 45.3 | 69.9 | 79.6 | 64.9 | <u>59.2</u> | 84.1 | 89.5 | 77.6 | <u>54.4</u> | 80.7 | 89.0 | 74.7 | <u>72.4</u> |
| | 25M | | 82 | <u>46.5</u> | <u>71.5</u> | <u>80.4</u> | <u>66.1</u> | **61.2** | **85.8** | **91.0** | **79.3** | **55.0** | **81.4** | **89.7** | **75.4** | **73.6** |

VindLU outperforms current SOTA by 7.8% on DiDeMo and 6.1% on ActivityNet

# Experiments

## Text-to-Video Retrieval

| Method | #PT | SSv2-label R1 | SSv2-label R5 | SSv2-template R1 | SSv2-template R5 | Avg |
|---|---|---|---|---|---|---|
| CLIP4Clip [49] | 400M | 43.1 | 71.4 | 77.0 | 96.6 | 77.9 |
| Singularity [31] | 17M | 47.4 | 75.9 | 77.6 | 96.0 | 80.0 |
| VINDLU | 5M | 51.2 | 78.8 | 82.2 | 98.9 | 82.7 |
| | 17M | 53.0 | 80.8 | **86.2** | 99.4 | **84.6** |
| | 25M | **53.1** | **81.8** | 83.3 | **100** | 84.4 |

VindLU outperforms SOTA by 5.7% and 8.6% on temporally-heavy SSv2-label and SSv2-template datasets.

# Experiments

## Video Question Answering

| Method | #PT | ANet | MSR-QA | MSR-MC | TVQA |
|---|---|---|---|---|---|
| ClipBERT [32] | 0.2M | - | 37.4 | 88.2 | - |
| ALPRO [34] | 5M | - | 42.1 | - | - |
| JustAsk [78] | 69M | 38.9 | 41.5 | - | - |
| VideoCLIP [75] | 136M | - | - | 92.1 | - |
| All-in-one [67] | 138M | - | 44.3 | 92.0 | - |
| MERLOT [86] | 180M | 41.4 | 43.1 | 90.9 | 78.7 |
| VIOLET [19] | 138M | - | 43.9 | 91.9 | - |
| Singularity [31] | 17M | 44.1 | 43.9 | 93.7 | - |
| OmniVL [69] | 17M | - | 44.1 | - | - |
| HERO [37] | 7.5M | - | - | - | 74.2 |
| FrozenBiLM [79] | 400M | 43.2 | 47.0 | - | 82.0 |
| VINDLU | 5M | 44.2 | 43.6 | 95.2 | **79.0** |
| | 17M | 44.6 | 43.8 | 96.7 | 78.8 |
| | 25M | **44.7** | **44.6** | **97.1** | **79.0** |

VindLU achieves competitive results across many VQA datasets.

# Experiments

## Action Recognition

| Method | TimeSformer [5] | OmniVL [69] | VINDLU |
|---|---|---|---|
| Top-1 acc. | 78.0 | 79.1 | **80.1** |

VindLU outperforms TimeSformer and OmniVL by 2.1% and 1.0% respectively.

# Conclusions

- We demystify the importance of various components used in VidL framework design.
- We provide a recipe for building a highly performant VidL model.
- Our model achieves SOTA performance on 9 video-language benchmarks.

**Check our code!**