

A. Implementation details

In this section, we provide additional implementation details of pretraining as well finetuning for both retrieval and classification tasks. We also include additional details regarding YT-VidLA-800M creation.

A.1. Additional pretraining details

For pretraining, we set the weight decay to 0.02. We use AdamW [54] optimizer and set β_1 and β_2 to 0.9 and 0.999 respectively. We set gradient clipping to 10.

A.2. Finetuning details for retrieval

For retrieval finetuning on all downstream datasets, we utilize 32 frames. For updating model parameters, we utilize AdamW optimizer with an initial learning rate of $1e - 5$ and decay it to $1e - 6$ using cosine annealing. We set the weight decay to 0.2. For VidLA-B/32 finetuning on all datasets, we set the batch size to 128, whereas, for VidLA-B/16 finetuning, we set the batchsize to 196. Unlike pretraining, we do not perform any multi-scale spatial crop for finetuning, rather, we perform center crop while preserving aspect ratio. We train VidLA-B/32 model for 440 steps on MSVD dataset. We respectively train VidLA-B/32 for $2.5\times$, $2.5\times$, $2\times$, and $6\times$ steps on MSR-VTT, DiDeMo, ActivityNet Captions, and VATEX datasets. For VidLA-B/model, we train for $4\times$, $1.5\times$, $2\times$, $1\times$, and $6\times$ steps on MSR-VTT, DiDeMo, ActivityNet Captions, MSVD and VATEX datasets, respectively.

A.3. Finetuning details for classification

For finetuning the video encoders from the pre-trained VidLA-B/32 for classification, we train for 100 epochs with a batch size of 512 for Kinetics-400 and 40 epochs with batch size 1024 for Something-Something-v2. For VidLA-B/16, we train for 75 epochs and 40 epochs respectively for Kinetics-400 and Something-Something-v2, each with a batch size of 512. We use the dense sampling approach of [24] for Kinetics-400, and TSN [84] uniform sampling for Something-Something-v2.

A.4. Additional details for YT-VidLA-800M creation

For caption generation, we use Blip-2 ViT-L OPT_{2.7B} [42]. We utilize Llama-2 13B Chat [78] model for caption and ASR subtitle summarization. We provide the summarization prompt in the following.

```
Summarize the following sentences
into a single sentence, not exceeding
25 words. Do not output any additional
text and use any external information.
<input text>
```

B. Zero-Shot Retrieval

We evaluate the zero-shot retrieval performance of VidLA-B/16 on the MSR-VTT dataset. We present the results in Table 8. We compare the performance with both two-tower and two-stage methods, without specific constraints on vision backbone size (e.g., ViT-B/16). We observe that VidLA outperforms all the prior works by a significant margin that employ similar-sized architectures. To be more specific, VidLA outperforms the recent two-stage method UMT [46] by 6%. Surprisingly, we also find that VidLA outperforms InternVideo [90] by 0.8%, even though InternVideo utilizes a larger vision backbone. These results provide further evidence for the effectiveness of VidLA in the video-language alignment task.

Method	R@1
Frozen [6]	18.7
VIOLET [26]	25.9
CLIP4Clip [55]	30.6
VINDLU† [12]	32.0
Singularity [38]	34.0
OmniVL [79]	34.6
VideoCoCa [95]	34.3
UMT-B† [46]	35.5
InternVideo(ViT-L)* [90]	40.7
VidLA-B/16	36.6
VidLA-B/16*	41.5

Table 8. Zero-shot retrieval results on MSR-VTT. * indicates inference with dual-softmax. † indicates two-stage method with candidate re-ranking.

C. Zero-Shot Multiple Choice

The objective of the zero-shot multiple choice task is to find the correct answer from a list of options. We evaluate the performance of VidLA-B/16 on the MSR-VTT dataset for this task and report the results in Table 9. We observe that VidLA outperforms the previous state-of-the-art, InternVideo, by 0.9%. We also notice that VidLA outperforms All-in-one by 11.9%. Overall, these results further demonstrate the generalizability of our large-scale video-language alignment training.

Method	Accuracy
All-in-one [80]	80.3
InternVideo(ViT-B) [90]	91.3
VidLA-B/16	92.2

Table 9. Zero-shot multiple choice results on MSR-VTT.

D. Zero-Shot Classification

We evaluate the performance of VidLA-B/16 on the zero-shot action classification task using the Kinetics-400

dataset. We only compare with methods that do not utilize Kinetics-400 data (labels and videos) for pretraining. To compare with prior art, we report results on both validation and test sets of Kinetics-400. We perform a ‘single-view’ evaluation and utilize uniformly sampled 32 frames. For evaluation, following CLIP [65], we obtain the embedding of all the class labels using the text encoder while utilizing the default prompt templates for Kinetics-400 classification. We consider the prediction as correct if the groundtruth label has the maximum similarity with the video embedding. We report the results in Table 10. We observe that our proposed method outperforms all prior works by a noticeable margin. To be precise, VidLA outperforms CLIP by around 11% on both validation and test sets. VidLA also outperforms, ActionCLIP [85] by 3.8%. We also notice that VidLA outperforms current state-of-the-art VFC [58] on both validation and test sets. Note that VFC [58] employs a verb focused contrastive loss during pretraining which greatly improves performance on action recognition task since the extracted verbs are, in principle, very similar to the downstream action recognition labels. Therefore, we also report a baseline score from VFC [58] which does not employ such a verb focused contrastive objective. VidLA outperforms the VFC-baseline by 3.5% on the Kinetics-400 test set. These results further validate the effectiveness of VidLA in action recognition, complementing the finetuned results reported in Table 7 of the main text.

Method	Top-1 Accuracy
Validation-Set	
CLIP [65]	48.9
ActionCLIP [85]	56.4
VFC [58]	59.4
VidLA-B/16	60.2
Test-Set	
Flamingo-3B [1]	45.2
Flamingo-80B [1]	49.1
Flamingo-9B [1]	49.7
CLIP [65]	47.9
VFC-Baseline [58]	55.6
VFC [58]	58.8
VidLA-B/16	59.1

Table 10. Zero-shot classification results on Kinetics-400.

E. Video-to-Text Retrieval Results

We present video-to-text retrieval results on DiDeMo, ActivityNet Captions, MSVD, and VATEX datasets in Table 11, 12, 13, and 14 respectively. Similar to the text-to-video retrieval results reported in the main text, we observe that VidLA outperforms prior art on video-to-text retrieval by a significant margin in most settings. However, InternVideo outperforms VidLA-B/16 on the VATEX dataset, likely due to its larger backbone architecture.

Method	DiDeMo Retrieval			
	R@1	R@5	R@10	Avg
CLIP-ViT-B/32				
CLIP4Clip [55]	42.5	70.6	80.2	64.4
CAMoE* [13]	45.5	71.2	—	—
DRL [87]	45.4	72.6	82.1	66.7
DiffusionRet* [35]	50.3	75.1	82.9	69.4
VidLA-B/32	55.0	82.8	88.9	75.6
VidLA-B/32*	62.2	85.1	91.1	79.5
CLIP-ViT-B/16				
DRL [87]	49.9	75.4	83.3	69.5
InternVideo(ViT-L)*	59.1	—	—	—
UMT†	59.5	84.9	90.5	78.3
VidLA-B/16	59.6	84.4	90.2	78.0
VidLA-B/16*	67.5	87.5	91.7	82.2

Table 11. Video-to-text retrieval on DiDeMo dataset. * indicates inference with dual-softmax. † indicates two-stage method with candidate re-ranking.

Method	ActivityNet Captions Retrieval			
	R@1	R@5	R@10	Avg
CLIP-ViT-B/32				
CLIP4Clip [55]	42.5	74.1	85.8	67.5
CenterCLIP [103]	44.5	75.7	86.2	68.8
DRL [87]	42.2	74.0	86.2	67.5
CAMoE* [13]	49.9	77.4	—	—
DiffusionRet* [35]	47.4	76.3	86.7	70.1
VidLA-B/32	60.8	85.0	91.8	79.2
VidLA-B/32*	69.1	88.8	93.7	83.9
CLIP-ViT-B/16				
DRL [87]	45.7	76.5	87.8	70.0
CenterCLIP [103]	46.7	77.1	88.0	70.6
UMT†	56.0	83.5	91.7	77.1
InternVideo(ViT-L)*	62.8	—	—	—
VidLA-B/16	63.4	87.5	93.2	81.4
VidLA-B/16*	72.9	90.6	94.9	86.1

Table 12. Video-to-text retrieval on ActivityNet Captions dataset. * indicates inference with dual-softmax. † indicates two-stage method with candidate re-ranking.

F. Effect of Image-Language Pretraining

A key objective of our work is to introduce minimal architectural modifications to effectively utilize pretrained image-language models. To analyze the effectiveness of image-language pretraining, we conduct experiments where we initialize the network weights randomly instead of using the CLIP weights. We evaluate the performance on three data scales and present the retrieval results on the MSR-VTT dataset in Figure 5. Figure 5 demonstrates that initializing the model weights with CLIP weights significantly outperforms the case where we initialize the model weights randomly at all data scales. This finding further validates the efficacy of VidLA in effectively utilizing the pretrained image-language weights and highlights the challenge of training from scratch for video-language alignment, which

Method	MSVD Retrieval			
	R@1	R@5	R@10	Avg
CLIP-ViT-B/32				
CLIP4Clip [55]	62.0	87.3	92.6	80.6
DRL [87]	62.3	86.3	92.2	80.3
CAMoE* [13]	49.8	79.2	87.0	72.0
DiffusionRet [35]	61.9	88.3	92.9	81.0
CenterCLIP [103]	63.5	86.4	92.6	80.8
VidLA-B/32	66.5	92.4	95.6	84.8
VidLA-B/32*	77.4	96.0	98.7	90.7
CLIP-ViT-B/16				
CenterCLIP [103]	68.4	90.1	95.0	84.5
DRL [87]	68.7	92.5	95.6	85.6
InternVideo(ViT-L)*	76.3	—	—	—
UMT†	74.0	94.6	97.3	88.6
VidLA-B/16	68.8	93.4	96.2	86.1
VidLA-B/16*	80.3	97.2	98.4	92.0

Table 13. Video-to-text retrieval on MSVD dataset. * indicates inference with dual-softmax. † indicates two-stage method with candidate re-ranking.

Method	VATEX Retrieval			
	R@1	R@5	R@10	Avg
CLIP-ViT-B/32				
DRL [87]	77.0	98.0	99.4	91.5
VidLA-B/32	79.8	99.5	99.8	93.0
VidLA-B/32*	85.0	99.8	99.9	94.9
CLIP-ViT-B/16				
DRL [87]	80.1	98.5	99.5	92.7
InternVideo(ViT-L)*	87.2	—	—	—
VidLA-B/16	81.1	99.5	99.9	93.5
VidLA-B/16*	85.6	99.8	100.0	95.1

Table 14. Video-to-text retrieval on VATEX dataset. * indicates inference with dual-softmax.

requires a large amount of training data.

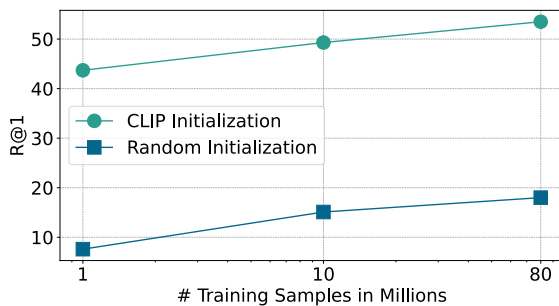


Figure 5. Effect of image-language pretraining on MSR-VTT.

G. Analysis of [MST] Token Configuration

To investigate the impact of various design aspects of [MST] tokens, we systematically vary the number of hierarchies, U , and the temporal scale, r and evaluate the performance

on the MSR-VTT dataset. We conduct these experiments on the 80 million data scale, consistent with most of the analysis done in the main text. We report the results in Table 15.

The first and the second rows of Table 15 demonstrate that incorporating [MST] tokens for a single hierarchy, $U = 1$, leads to noticeable performance gains. The third and fourth rows further illustrate that increasing the number of hierarchies to 2 provides additional improvements, and setting the temporal scale to a higher value leads to better results. Finally, the last two rows indicate that adding one more hierarchy provides further improvement. However, we do not notice a significant difference between different temporal scales. Note that the [MST] token design in the last row is our default setting. The results in Table 15 provides further support for our hypothesis that incorporating hierarchy into the design of [MST] tokens plays a crucial role in improving video-language alignment performance.

$U-V-r$	MSR-VTT Retrieval			
	R@1	R@5	R@10	Avg
\times	49.1	75.3	83.5	69.3
1-4-1	51.3	76.5	85.0	70.9
2-4-2	52.1	76.6	84.2	71.0
2-4-3	52.3	76.9	85.7	71.6
3-4-3	53.2	77.0	85.6	71.9
3-4-2	53.5	77.5	85.6	72.2

Table 15. Retrieval performance on MSR-VTT with different [MST] designs.

H. Results on Hierarchical Dataset

We demonstrated VidLA’s efficacy on both short and long videos in Figure 4 (left). To further illustrate VidLA’s effectiveness in a strictly hierarchical setting, we constructed a temporally hierarchical test set with 1K short, mid and long clip triplets centered around the same timestamp. Zero-shot retrieval performance in Table 16 shows VidLA outperforms current state-of-the-art method, CLIPViP, even when trained on the same pretraining dataset (YT-VidLA-80M).

Method	Short	Medium	Long
CLIPViP-B/32	85.9	86.8	88.1
VidLA-B/32	86.8	90.5	92.5

Table 16. Text2Video retrieval R@1 on a 1k hierarchical test sets

I. Analysis of Pretraining Dataset

To analyze the effect of different pretraining datasets, we pretrain VidLA-B/32 on different pretraining datasets and later finetune on downstream datasets and report the results in Table 17. The first row reports the results where no pretraining on video data was performed and the model was only finetuned on the downstream dataset. From second

and third row we notice that training on YT-VidLA-10M outperforms training on WebVid-10M. The last two rows demonstrate the advantage of our large-scale training where the performance improves significantly with scale.

Dataset	MSR-VTT	DiDeMo	ActivityNet
WIT-400M	42.9	40.7	40.7
WebVid-10M	43.6	46.6	46.9
YT-VidLA-10M	49.3	46.3	49.7
YT-VidLA-80M	53.5	52.3	53.9
YT-VidLA-800M	55.6	56.9	61.3

Table 17. Text-to-Video retrieval R@1 with different pretraining datasets