Frozen in Time:
A Joint Video and Image Encoder for End-to-End Retrieval

ICCV 2021

Authors: Max Bain, Arsha Nagrani, Gul Varol, Andrew Zisserman

Presenters: Ziyang Wang, Han Wang, Han Lin

Pretraining Datasets

- **Task**: text-to-video retrieval

- **Challenge**:
  - Available large scale video-text training datasets (e.g., HowTo100M) are noisy.
  - Competitive performance is achieved only at scale through large amounts of compute.

- **Main contributions of this paper:**
  - An end-to-end trainable model designed to take advantage of both large-scale image and video captioning datasets
  - A new video-text pre-training dataset WebVid-2M, comprised of over two million videos with weak captions scraped from the internet.

Architecture Details

- **Video Encoder**
  - **Patch embedding layer**
    - 2D convolutional layer with N target patch of size P=16, and output channels d=768

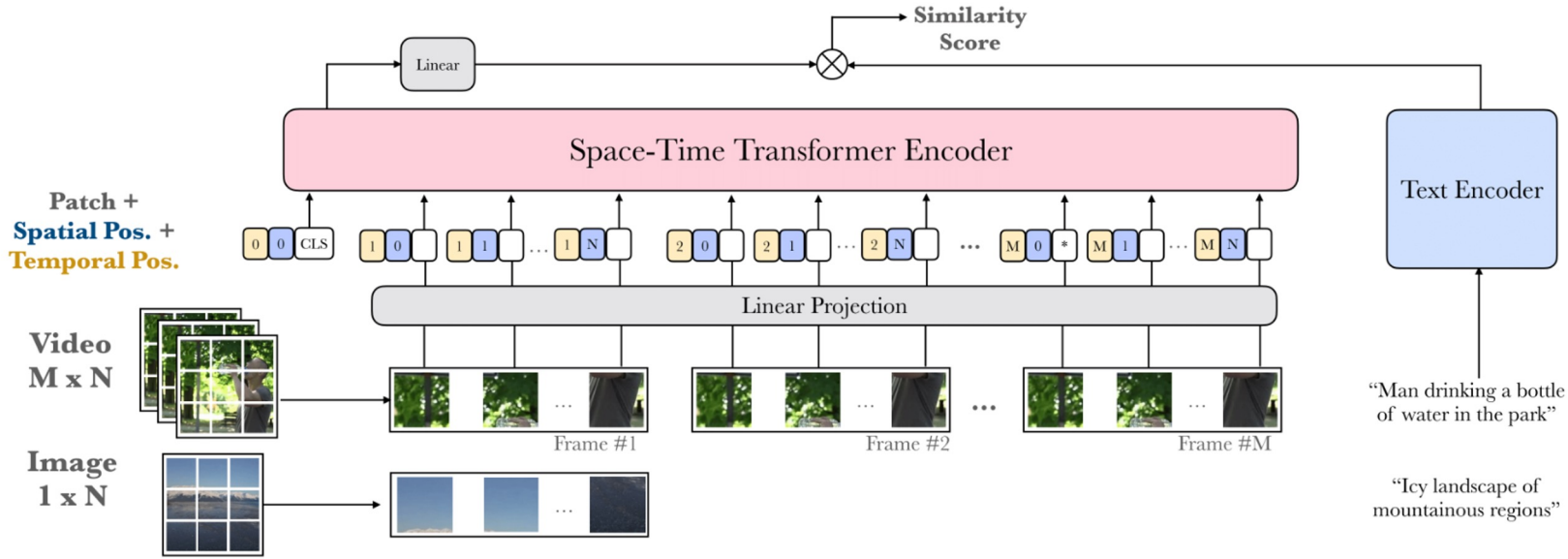


Figure 1: **Joint Image and Video Training:** Our dual encoding model consists of a visual encoder for images and video and a text encoder for captions. Unlike 2D or 3D CNNs, our space-time transformer encoder allows us to train flexibly on both images and videos with captions jointly, by treating an image as a single frame video.

Architecture Details

- **Video Encoder**
  - **Learnable positional space and time embeddings**
    - Positional space embedding: M * d, where M is max number of input video frames
    - Positional time embedding: N * d, where N is max number of patches of size P
  - **[CLS] embedding:** 1 * d, which is used to produce the final video embedding

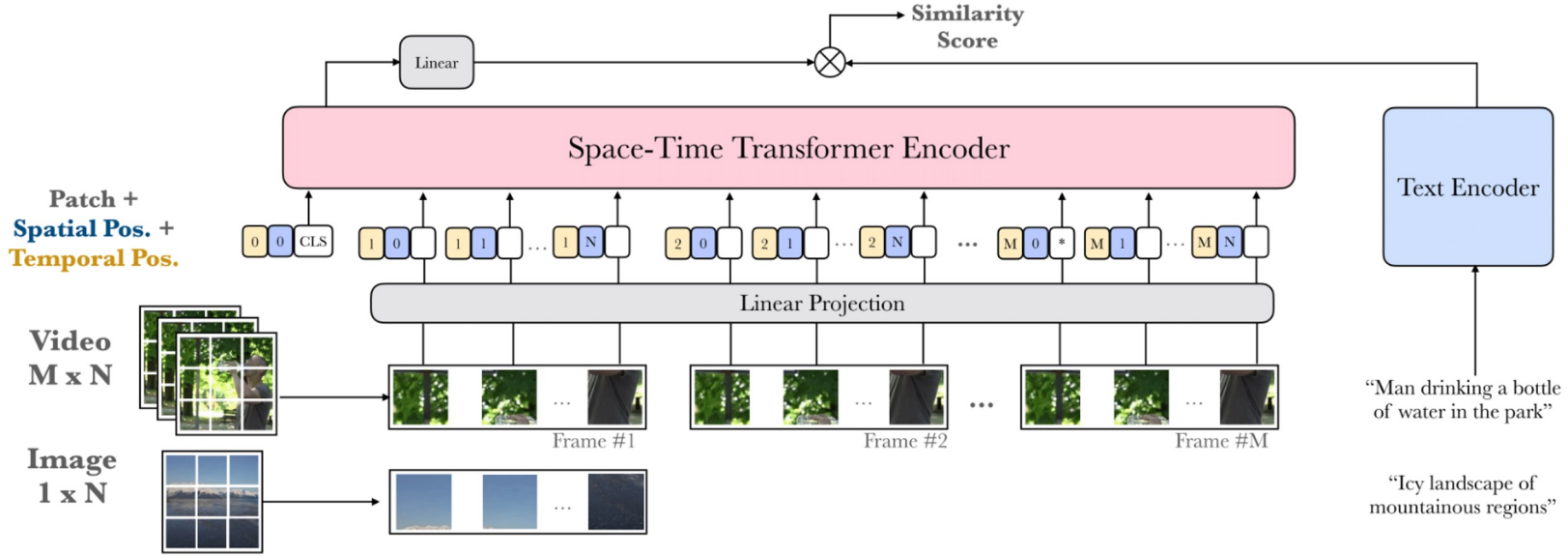


Figure 1: **Joint Image and Video Training:** Our dual encoding model consists of a visual encoder for images and video and a text encoder for captions. Unlike 2D or 3D CNNs, our space-time transformer encoder allows us to train flexibly on both images and videos with captions jointly, by treating an image as a single frame video.

Architecture Details

- **Video Encoder**
  - **Space-time attention block**
    - A stack of 12 space-time attention blocks

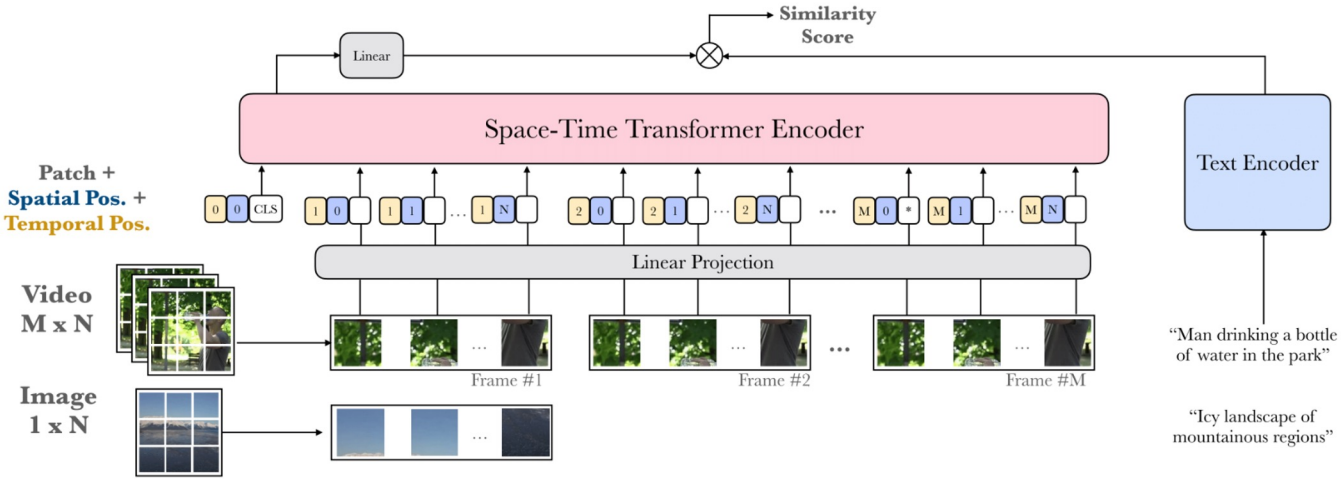
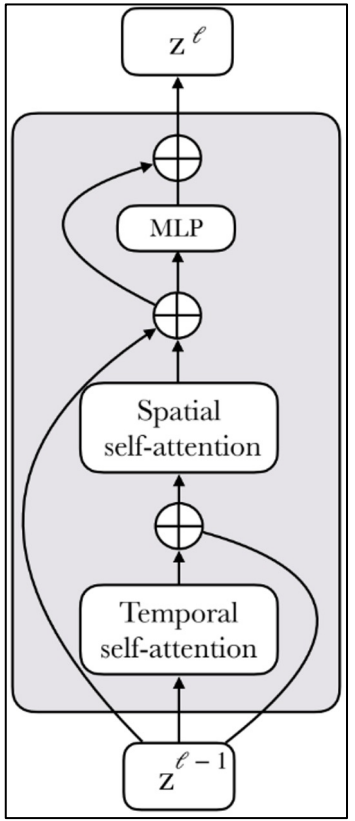


Figure 1: **Joint Image and Video Training:** Our dual encoding model consists of a visual encoder for images and video and a text encoder for captions. Unlike 2D or 3D CNNs, our space-time transformer encoder allows us to train flexibly on both images and videos with captions jointly, by treating an image as a single frame video.



Architecture Details

- **Text Encoder**
  - Initialized as distilbert-base-uncased
- **Linear Projection Layer**
  - Projects text and video embeddings to a common space

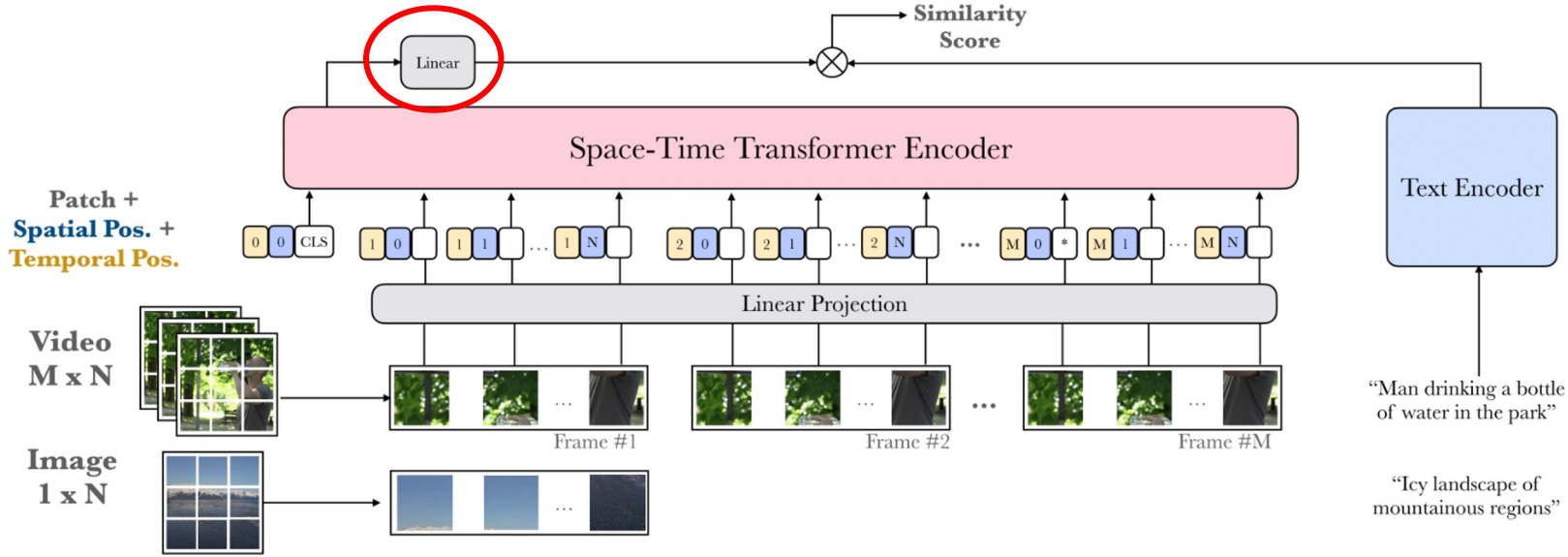


Figure 1: **Joint Image and Video Training:** Our dual encoding model consists of a visual encoder for images and video and a text encoder for captions. Unlike 2D or 3D CNNs, our space-time transformer encoder allows us to train flexibly on both images and videos with captions jointly, by treating an image as a single frame video.

Architecture Details

- **Loss**
  - **Positive pairs**: Matched text-video pairs in a batch
  - **Negative pairs:** All other pairwise combinations in a batch

- Scalable to large scale retrieval at inference time

$$L_{v2t} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{B}\exp(x_i^\top y_j/\sigma)}$$

$$L_{t2v} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{B}\exp(y_i^\top x_j/\sigma)}$$

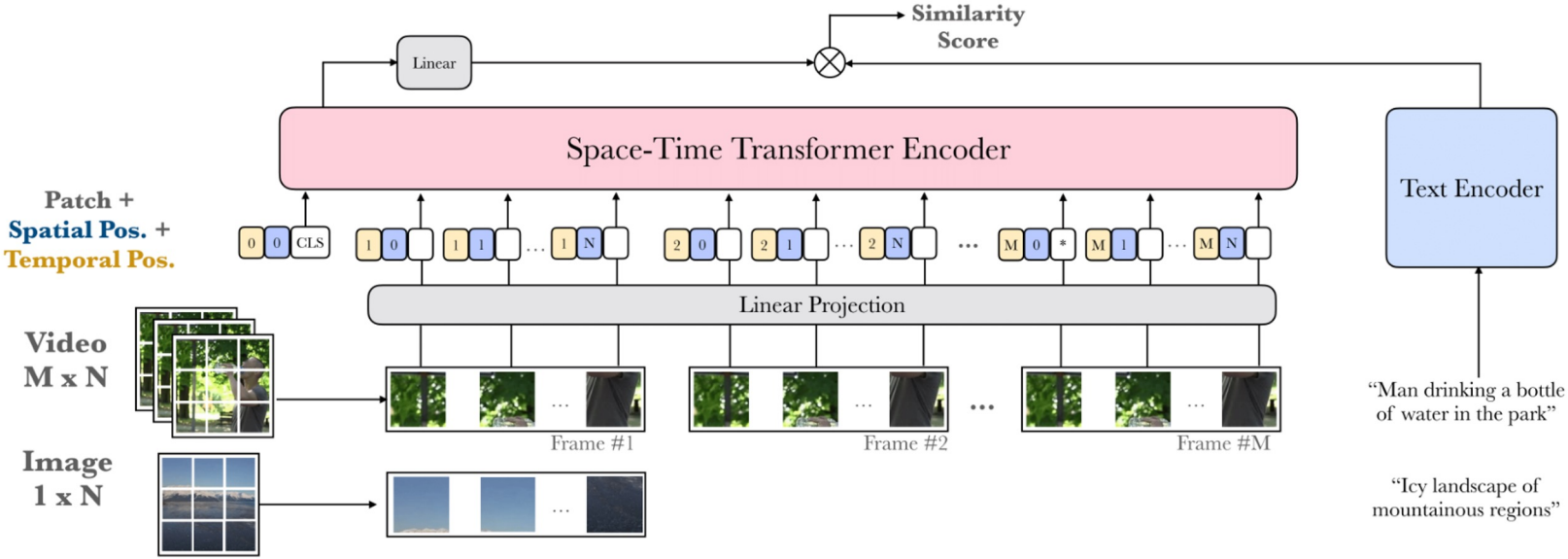


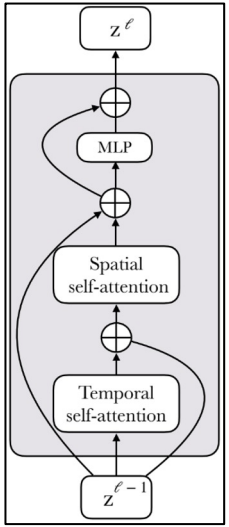
Figure 1: **Joint Image and Video Training:** Our dual encoding model consists of a visual encoder for images and video and a text encoder for captions. Unlike 2D or 3D CNNs, our space-time transformer encoder allows us to train flexibly on both images and videos with captions jointly, by treating an image as a single frame video.

Pretraining Datasets

- **Video:** WebVid2M
  - Manually generated captions
- **Image:** Google Conceptual Captions
  - ~3.3M image-text pairs

dataset	domain	#clips	avg dur. (secs)	#sent	time (hrs)
MPII Cook [54]	cooking	44	600	6K	8
TACos [51]	cooking	7K	360	18K	15.9
DideMo [3]	flickr	27K	28	41K	87
MSR-VTT [72]	youtube	10K	15	200K	40
Charades [60]	home	10K	30	16K	82
LSMDC15 [53]	movies	118K	4.8	118K	158
YouCook II [78]	cooking	14K	316	14K	176
ActivityNet [29]	youtube	100K	180	100K	849
CMD [5]	movies	34K	132	34K	1.3K
WebVid-2M	open	**2.5M**	18	**2.5M**	**13K**
HT100M [44]	instruction	136M	4	136M	134.5K

- **Joint image-video training:**
  - Alternating batches between the image and video datasets
- **Weight initialization:**
  - Initialize the spatial attention weights with ViT weights trained on ImageNet-21k, and initialize the temporal attention weights to zero.
- **Temporal curriculum training:**
  - The space-time transformer allows a variable number of input video frames.
  - Initially training on fewer frames can drastically save computation



Effect of pretraining

1) trained from scratch
2) initialised with ImageNet weights, and then finetuned
3) initialised with ImageNet, and then pretrained on different visual-text datasets before finetuning

	Pre-training	#pairs	R@1	R@10	MedR
1)	-	-	5.6	22.3	55
2)	ImageNet		15.2	54.4	9.0
3)	HowTo-17M subset	17.1M	24.1	63.9	5.0
3)	CC3M	3.0M	24.5	62.7	5.0
3)	WebVid2M	2.5M	26.0	64.9	5.0
3)	**CC3M + WebVid2M**	5.5M	**27.3**	**68.1**	**4.0**

Performance on MSR-VTT

Pretraining sources extended

- Our proposed method can achieve reasonable performance on downstream video data with a small number of text-image pairs pretraining alone.
- Increasing the number of pretraining pairs consistently improves downstream performance.

Pre-training	#pairs	R@1	R@5	R@10	MedR
COCO	0.6M	27.2	56.1	67.5	4.0
WV-2M	2.5M	27.5	56.6	67.6	4.0
WV-10M	10M	28.9	57.2	68.6	4.0
CC3M, WV2M	5.0M	31.0	59.5	70.5	3.0
CC3M, WV2M, COCO	5.6M	32.5	61.5	71.2	3.0
CC3M, WV10M	13.0M	33.4	59.2	70.7	3.0
CC3M, CC12M, WV10M	25.0M	34.0	61.4	73.1	3.0

Performance on MSR-VTT

Effect of #frames and curriculum learning

- Performing the temporal expansion at pretraining stage is better than at finetuning

PT #frames	FT #frames	R@1	R@10	MedR	PTT (hrs)
1	1	18.8	56.6	7.0	16.2
1	4	24.9	67.1	5.0	16.2
4	4	26.0	64.9	5.0	45.6
1⇒4	4	26.6	65.5	5.0	22.1
8	8	25.4	67.3	4.0	98.0
1⇒4⇒8	8	27.4	67.3	4.0	36.0

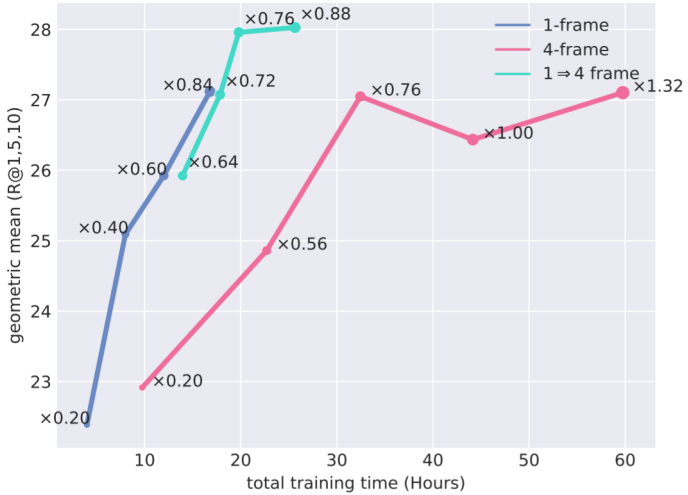
⇒ indicates a within-dataset curriculum learning strategy

Effect of #frames and curriculum learning

- Curriculum can obtain similar or better performance with much less computational cost.

PT #frames	FT #frames	R@1	R@10	MedR	PTT (hrs)
1	1	18.8	56.6	7.0	16.2
1	4	24.9	67.1	5.0	16.2
4	4	26.0	64.9	5.0	45.6
1⇒4	4	26.6	65.5	5.0	22.1
8	8	25.4	67.3	4.0	98.0
1⇒4⇒8	8	27.4	67.3	4.0	36.0

⇒ indicates a within-dataset curriculum learning strategy



× denotes the multiple of dataset epochs completed

Comparison to the State of the Art

Results on MSR-VTT (text-to-video retrieval)

Method	E2E†	Vis Enc. Init.	Visual-Text PT	#pairs PT	R@1	R@5	R@10	MedR
JSFusion [75]	✓	-	-	-	10.2	31.2	43.2	13.0
HT MIL-NCE [44]	✓	-	HowTo100M	136M	14.9	40.2	52.8	9.0
ActBERT [80]	✓	VisGenome	HowTo100M	136M	16.3	42.8	56.9	10.0
HERO [34]	✓	ImageNet, Kinetics	HowTo100M	136M	16.8	43.4	57.7	-
VidTranslate [28]	✓	IG65M	HowTo100M	136M	14.7	-	52.8	
NoiseEst. [2]	✗	ImageNet, Kinetics	HowTo100M	136M	17.4	41.6	53.6	8.0
CE [38]	✗	Numerous experts†	-		20.9	48.8	62.4	6.0
UniVL [40]	✗	-	HowTo100M	136M	21.2	49.6	63.1	6.0
ClipBERT [32]	✓	-	COCO, VisGenome	5.6M	22.0	46.8	59.9	6.0
AVLnet [55]	✗	ImageNet, Kinetics	HowTo100M	136M	27.1	55.6	66.6	4.0
MMT [21]	✗	Numerous experts†	HowTo100M	136M	26.6	57.1	69.6	4.0
T2VLAD [69]	✗	Numerous experts†	-		29.5	59.0	70.1	4.0
Support Set [48]	✗	IG65M, ImageNet	-	-	27.4	56.3	67.7	3.0
Support Set [48]	✗	IG65M, ImageNet	HowTo100M	136M	30.1	58.5	69.3	**3.0**
Ours	✓	ImageNet	CC3M	3M	25.5	54.5	66.1	4.0
Ours	✓	ImageNet	CC3M,WV-2M	5.5M	**31.0**	**59.5**	**70.5**	**3.0**
Ours	✓	ImageNet	CC3M,WV-2M,COCO	6.1M	**32.5**	**61.5**	**71.2**	**3.0**
Zero-shot								
HT MIL-NCE [44]	✓	-	HowTo100M	136M	7.5	21.2	29.6	38.0
SupportSet [48]		IG65M, ImageNet	HowTo100M	136M	8.7	23.0	31.1	31.0
Ours	✓	ImageNet	CC3M,WV-2M	5.5M	**23.2**	**44.6**	**56.6**	**7.0**
Ours	✓	ImageNet	CC3M,WV-2M,COCO	6.1M	**24.7**	**46.9**	**57.2**	**7.0**

Comparison to the State of the Art

Results on MSVD

Method	R@1	R@5	R@10	MedR
VSE [27]	12.3	30.1	42.3	14.0
VSE++ [20]	15.4	39.6	53.0	9.0
Multi. Cues [45]	20.3	47.8	61.1	6.0
CE [38]	19.8	49.0	63.8	6.0
Support Set [48]	23.0	52.8	65.8	5.0
Support Set [48] (HowTo PT)	28.4	60.0	72.9	4.0
Ours	**33.7**	**64.7**	**76.3**	**3.0**

Results on DiDeMo

Method	GT prop.	R@1	R@5	R@10	MedR
S2VT [64]		11.9	33.6	-	13.0
FSE [77]		13.9	36.0	-	11.0
CE [38]		16.1	41.1	-	8.3
ClipBERT [32]	✓	20.4	44.5	56.7	7.0
Ours		**31.0**	**59.8**	**72.4**	**3.0**
Ours	✓	**34.6**	**65.0**	**74.7**	**3.0**
Zero-shot					
Ours		21.1	46.0	56.2	7.0
Ours	✓	20.2	46.4	58.5	7.0

Comparison to the State of the Art

Results on LSMDC

Method	R@1	R@5	R@10	MedR
JSFusion [75]	9.1	21.2	34.1	36.0
MEE [43]	9.3	25.1	33.4	27.0
CE [38]	11.2	26.9	34.8	25.3
MMT (HowTo100M) [21]	12.9	29.9	40.1	**19.3**
Ours	**15.0**	**30.8**	**40.3**	20.0

Results on Flickr30K (Text-to-**image** retrieval)

Method	Vis PT. size	R@1	R@5	R@10
SCANM [31]	VisGenObj (3.8M)	48.6	77.7	85.2
IMRAM [11]	VisGenObj (3.8M)	53.9	79.4	87.2
SGRAF [18]	VisGenObj (3.8M)	58.5	83.0	88.8
Ours	CC (3.0M)	54.2	83.2	89.8
Ours	CC,WV-2M (5.5M)	61.0	87.5	92.7

Comparison to the State of the Art

Results on LSMDC

Method	R@1	R@5	R@10	MedR
JSFusion [75]	9.1	21.2	34.1	36.0
MEE [43]	9.3	25.1	33.4	27.0
CE [38]	11.2	26.9	34.8	25.3
MMT (HowTo100M) [21]	12.9	29.9	40.1	**19.3**
Ours	**15.0**	**30.8**	**40.3**	20.0

Results on Flickr30K (Text-to-**image** retrieval)

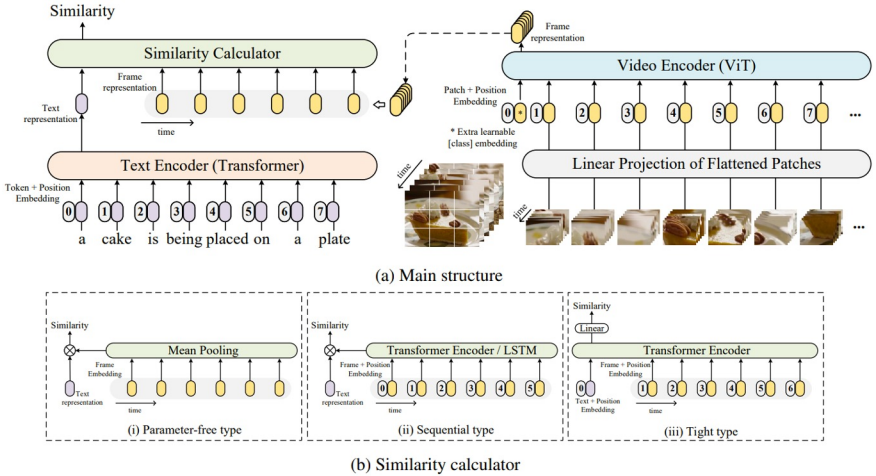
Method	Vis PT. size	R@1	R@5	R@10
SCANM [31]	VisGenObj (3.8M)	48.6	77.7	85.2
IMRAM [11]	VisGenObj (3.8M)	53.9	79.4	87.2
SGRAF [18]	VisGenObj (3.8M)	58.5	83.0	88.8
Ours	CC (3.0M)	54.2	83.2	89.8
Ours	CC,WV-2M (5.5M)	61.0	87.5	92.7

Advantages of Frozen in Time

1) Unified framework on visual information **VS** video-only
2) Collect WebVid2M dataset with clean caption **VS** train from noisy data
3) Inspiration for future works (case study of ICCV23 video-text retrieval papers)
    a) MIL-NCE (1/7)
    b) Image-text learning to video-text learning (7/7)

Methods	R@1	R@5	R@10	MdR↓	MnR↓
CLIP4Clip [39]	47.1	74.1	81.8	2.0	14.9
TI (Token-Wise)	48.4	74.2	83.3	2.0	14.1
+ DSA	49.6	75.5	84.9	2.0	12.5
+ DUA†	50.1	75.8	84.6	1.5	12.8
+ KL† (UATVR)	**50.8**	**76.3**	**85.5**	**1.0**	**12.4**
+ DUA*	50.0	75.8	83.9	1.5	12.9
+ KL*	50.6	75.9	84.9	**1.0**	12.8

Table 1. Ablation study of different components. † denotes the implementation with MIL-NCE contrast and * is implemented with soft contrastive loss via Monte-Carlo estimation [45].



[1] UATVR: Uncertainty-Adaptive Text-Video Retrieval, Fang et al. ICCV23
[2] CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval, Luo et al.