vid-TLDR: Training Free Token merging for Light-weight Video Transformer

Supplementary Material

A. A deeper exploration into temporal bias

We here present the discussion for the behavior of video Transformers regarding temporal bias. We have observed that the positional encoding in temporal-spatial attention even has learned the data bias, *i.e.*, the main contents of the clip are generally positioned at the front or middle. For better understanding, we visualize the attention map built with 1) positional encoding (PE) + representation, 2) representation only, and 3) PE only. Figure 1 reveals that PE of video Transformers causes the temporal bias, *i.e.* focusing on the early frame.

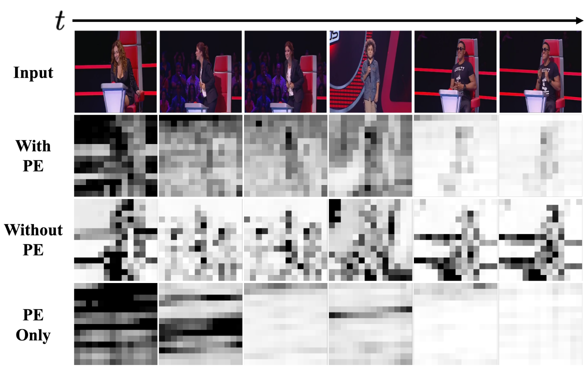

Figure 1. Visualization of temporal bias caused by PE.

B. Detailed reduced number of tokens

In this section, we present the hyperparameters indicating the number of reduced tokens for each dataset and task, as shown in Table A. According to the main objective of vid-TLDR, which focuses on early-stage merging, all experiments involved reduction in the first four layers.

Table A. **Number of tokens reduced in each experiment.**

Task	Method	Dataset	Block index				
			1	2	3	4	5 - Top
Retrieval	UMT-B	MSRVTT	600	200	100	0	0
		MSVD	450	250	200	0	0
		ActivityNet	500	0	0	0	0
		DiDeMo	250	150	150	150	0
		LSMDC	400	150	150	0	0
		SSV2-Label	200	200	200	100	0
		SSV2-Template	400	300	0	0	0
	UMT-L	MSRVTT	550	400	0	0	0
		MSVD	600	200	100	100	0
		ActivityNet	600	300	0	0	0
		DiDeMo	800	100	0	0	0
		LSMDC	250	250	200	200	0
		SSV2-Label	450	250	100	0	0
		SSV2-Template	350	300	100	50	0
VQA	UMT-B	MSRVTT	600	200	0	0	0
		MSVD	600	200	0	0	0
	UMT-L	MSRVTT	400	400	100	0	0
		MSVD	400	400	100	0	0

C. Detailed video-text retrieval

In Table 2 of the main paper, we provide video-text retrieval by averaging the performances of video-to-text retrieval and text-to-retrieval. Here, we provide the detailed results of each task in Table B with MSRVTT [8], MSVD [4], ActivityNet [3], DiDeMo [1], LSMDC [7], SSV2-Label/Template [5].

D. Zero-shot retrieval results

To validate the effectiveness of vid-TLDR, we report additional zero-shot retrieval results for baseline model [6], existing token merging method (ToMe [2]) and ours across MSVD [4], MSRVTT [8], DiDeMo [1], ActivityNet [3]. All experiments were conducted with evaluation only, without additional training. The results for text-to-video and video-to-text retrieval accuracy across these four datasets are presented in Table C, D, E, F, and G. We conducted a grid search to determine the effective number of tokens to reduce. In the case of ToMe, tokens were merged uniformly across layers following the default configuration, and a grid search was conducted with 20-25 tokens per layer. On the other hand, our method merged in the first four front layers, exploring total reductions of 300, 330, and 360 tokens. Across all configurations, vid-TLDR consistently achieves higher R@1 with lower FLOPs compared to ToMe.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2022. 1

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1

[4] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th An-*

Table B. **Detailed results on text-to-video & video-to-text retrieval.**

Dataset	Metric	Type	UMT-B			UMT-L		
			Base	ToMe	Ours	Base	ToMe	Ours
MSRVTT	GFLOPs ↓	(G)	303.3	231.4	178.0	984.6	529.7	563.1
	R@1 ↑	T2V	51.0	46.6	50.9	58.8	55.6	58.1
		V2T	49.0	47.4	50.7	58.6	56.0	58.7
	R@5 ↑	T2V	76.5	71.8	75.8	81.0	79.7	81.0
		V2T	77.0	74.6	75.5	81.6	79.4	81.6
	R@10 ↑	T2V	84.2	80.9	83.9	87.1	85.3	86.8
		V2T	84.7	83.2	83.9	86.5	86.8	86.9
MSVD	GFLOPs ↓	(G)	303.3	218.7	181.3	984.6	574.5	563.1
	R@1 ↑	T2V	50.8	47.8	50.5	58.2	57.1	57.9
		V2T	73.3	71.3	74.9	82.4	81.9	82.7
	R@5 ↑	T2V	79.7	78.2	79.7	83.9	83.6	83.8
		V2T	88.8	89.4	90.0	94.6	93.9	94.5
	R@10 ↑	T2V	85.7	85.5	86.3	89.6	89.4	89.4
		V2T	93.7	92.5	93.3	96.7	96.0	96.3
ActivityNet	GFLOPs ↓	(G)	303.3	236.8	227.6	984.6	574.5	572.9
	R@1 ↑	T2V	58.3	51.9	57.8	66.8	63.3	66.7
		V2T	56.0	51.4	55.4	64.4	61.7	63.9
	R@5 ↑	T2V	83.9	80.8	83.7	89.1	87.3	88.6
		V2T	83.5	80.6	83.2	89.1	86.5	88.7
	R@10 ↑	T2V	91.5	89.4	91.5	94.9	93.6	94.4
		V2T	91.7	89.9	91.1	94.8	93.6	94.5
DiDeMo	GFLOPs ↓	(G)	303.3	218.7	212.8	984.6	574.5	559.0
	R@1 ↑	T2V	63.7	56.4	64.1	72.5	68.7	72.3
		V2T	60.5	58.2	60.8	69.2	67.3	68.5
	R@5 ↑	T2V	87.8	82.3	87.2	91.2	89.5	91.2
		V2T	85.7	82.8	85.1	89.9	89.2	89.8
	R@10 ↑	T2V	93.5	89.0	92.7	94.8	94.6	94.2
		V2T	90.7	89.5	90.4	94.1	92.9	93.8
LSMDC	GFLOPs ↓	(G)	303.3	223.2	206.2	984.6	574.5	583.7
	R@1 ↑	T2V	32.7	25.6	32.9	43.0	39.7	43.1
		V2T	32.7	29.0	31.8	41.4	38.7	40.7
	R@5 ↑	T2V	54.7	47.6	53.6	65.5	61.8	64.5
		V2T	53.5	50.5	52.9	64.3	61.4	63.6
	R@10 ↑	T2V	63.4	57.6	63.3	73.0	61.8	71.4
		V2T	63.2	57.0	63.1	71.5	67.9	70.2
SSV2-label	GFLOPs ↓	(G)	303.3	232.2	212.9	984.6	627.2	610.9
	R@1 ↑	T2V	64.1	58.7	64.3	73.3	69.4	73.1
		V2T	63.8	61.7	63.3	71.5	70.3	71.1
	R@5 ↑	T2V	88.2	84.9	88.0	93.7	92.1	93.3
		V2T	88.3	87.2	87.4	93.1	92.3	92.7
	R@10 ↑	T2V	92.7	90.8	92.6	96.6	95.8	96.6
		V2T	93.2	92.4	92.9	96.7	95.8	96.4
SSV2-Template	GFLOPs ↓	(G)	303.3	241.4	203.7	984.6	627.2	572.9
	R@1 ↑	T2V	87.9	85.1	87.4	90.8	88.5	90.2
		V2T	61.3	59.5	60.6	66.1	65.5	66.0
	R@5 ↑	T2V	99.4	100.0	99.4	99.4	98.9	100.0
		V2T	88.5	87.9	87.4	92.3	91.4	91.7
	R@10 ↑	T2V	100.0	100.0	100.0	100.0	100.0	100.0
		V2T	93.7	93.1	92.6	95.6	95.5	95.8

Table C. **Comparison of zero-shot retrieval performance for four datasets**.

Dataset	#Pairs	Metic	UMT-B Base	UMT-B ToMe	UMT-B Ours	UMT-L Base	UMT-L ToMe	UMT-L Ours
MSVD	5M	GFLOPs ↓	78.5	66.2	**50.6**	267.8	182.1	**165.3**
		R@1 ↑	47.4	45.6 (-1.8)	**48.1** (+0.7)	55.3	53.6 (-1.7)	**56.3** (+1.0)
		R@5 ↑	72.2	70.8 (-1.4)	**72.5** (+0.3)	79.4	78.9 (-0.5)	**79.6** (+0.2)
		R@10 ↑	80.2	79.8 (-0.4)	**80.4** (+0.2)	85.9	85.1 (-0.8)	**86.1** (+0.2)
	17M	GFLOPs ↓	78.5	62.8	**50.2**	267.8	182.1	**165.3**
		R@1 ↑	52.0	50.7 (-1.3)	**53.3** (+1.3)	**62.7**	61.0 (-1.7)	**62.7** (±0.0)
		R@5 ↑	75.7	74.8 (-0.9)	**77.0** (+1.3)	83.7	83.2 (-0.5)	**83.8** (+0.1)
		R@10 ↑	83.6	83.0 (-0.6)	**84.0** (+0.4)	89.7	88.8 (-0.9)	**90.3** (+0.6)
	25M	GFLOPs ↓	78.5	66.2	**50.6**	267.8	182.1	**153.6**
		R@1 ↑	52.1	50.1 (-2.0)	**53.6** (+1.5)	61.8	60.3 (-1.5)	**62.5** (+0.7)
		R@5 ↑	77.1	76.1 (-1.0)	**78.1** (+1.0)	83.3	82.5 (-0.8)	**83.2** (-0.1)
		R@10 ↑	84.7	84.1 (-0.6)	**85.4** (+0.7)	88.8	88.3 (-0.5)	**88.8** (±0.0)
MSRVTT	5M	GFLOPs ↓	78.5	66.2	**50.2**	267.8	182.1	**164.8**
		R@1 ↑	27.9	26.4 (-1.5)	**27.7** (-0.2)	31.8	30.4 (-1.4)	**32.7** (+0.9)
		R@5 ↑	49.8	46.7 (-3.1)	**49.8** (±0.0)	54.7	52.4 (-2.3)	**55.1** (+0.4)
		R@10 ↑	58.4	55.9 (-2.5)	**58.3** (-0.1)	64.2	62.9 (-1.3)	**64.6** (+0.4)
	17M	GFLOPs ↓	78.5	66.2	**50.6**	267.8	182.1	**163.0**
		R@1 ↑	33.6	32.5 (-1.1)	**33.8** (+0.2)	40.6	37.6 (-3.0)	**39.9** (-0.7)
		R@5 ↑	56.4	54.1 (-2.3)	**57.0** (+0.6)	62.0	60.1 (-1.9)	**61.9** (-0.1)
		R@10 ↑	66.4	63.9 (-2.5)	**66.6** (+0.2)	71.4	68.9 (-2.5)	**70.9** (-0.5)
	25M	GFLOPs ↓	78.5	66.2	**44.7**	267.8	160.9	**146.5**
		R@1 ↑	32.8	31.1 (-1.7)	**33.3** (+0.5)	38.9	36.3 (-2.6)	**39.2** (+0.3)
		R@5 ↑	54.3	52.8 (-1.5)	**54.7** (+0.4)	61.1	59.0 (-2.1)	**61.2** (+0.1)
		R@10 ↑	32.8	31.1 (-1.7)	**33.3** (+0.5)	38.9	36.3 (-2.6)	**39.2** (+0.3)
DIDEMO	5M	GFLOPs ↓	78.5	66.2	**47.3**	267.8	182.1	**166.2**
		R@1 ↑	35.8	29.7 (-6.1)	**35.9** (+0.1)	35.1	33.7 (-1.4)	**36.6** (+1.5)
		R@5 ↑	61.9	55.5 (-6.4)	**60.6** (-1.3)	60.2	60.0 (-0.2)	**62.1** (+1.9)
		R@10 ↑	70.1	65.1 (-5.0)	**69.0** (-1.1)	68.7	69.5 (+0.8)	**71.0** (+2.3)
	17M	GFLOPs ↓	78.5	62.8	**44.5**	267.8	182.1	**166.7**
		R@1 ↑	41.1	38.6 (-2.5)	**42.3** (+1.2)	46.5	45.4 (-1.0)	**48.7** (+2.2)
		R@5 ↑	66.7	64.7 (-2.0)	**68.4** (+1.7)	71.1	71.8 (+0.7)	**74.3** (+3.2)
		R@10 ↑	75.4	73.8 (-1.6)	**77.0** (+1.6)	79.2	80.1 (+0.9)	**81.3** (+2.1)
	25M	GFLOPs ↓	78.5	62.8	**48.2**	267.8	182.1	**162.5**
		R@1 ↑	41.0	38.6 (-2.4)	**44.0** (+3.0)	49.3	47.9 (-1.3)	**52.0** (+2.8)
		R@5 ↑	66.6	64.5 (-2.1)	**68.9** (+2.3)	73.9	72.6 (-1.3)	**74.9** (+1.1)
		R@10 ↑	75.8	73.0 (-2.8)	**77.7** (+1.9)	80.2	81.1 (+0.9)	**82.4** (+2.2)
ANET	5M	GFLOPs ↓	78.5	66.2	**49.0**	267.8	182.1	**162.0**
		R@1 ↑	27.1	24.9 (-2.2)	**28.1** (+1.0)	31.0	29.7 (-1.3)	**31.3** (+0.3)
		R@5 ↑	51.6	49.1 (-2.5)	**52.4** (+0.8)	59.7	57.9 (-1.8)	**59.6** (-0.1)
		R@10 ↑	63.0	61.0 (-2.0)	**64.1** (+1.1)	71.7	70.2 (-1.5)	**71.6** (-0.1)
	17M	GFLOPs ↓	78.5	62.8	**45.0**	267.8	182.1	**151.7**
		R@1 ↑	32.7	31.7 (-1.0)	**33.9** (+1.2)	41.8	40.3 (-1.5)	**41.7** (-0.1)
		R@5 ↑	57.7	57.4 (-0.3)	**59.6** (+1.9)	68.6	67.1 (-1.5)	**68.8** (+0.2)
		R@10 ↑	69.2	68.9 (-0.3)	**70.7** (+1.5)	79.2	78.2 (-1.0)	**79.1** (-0.1)
	25M	GFLOPs ↓	78.5	62.8	**51.0**	267.8	182.1	**152.7**
		R@1 ↑	34.2	32.6 (-1.6)	**35.7** (+1.5)	40.7	41.0 (+0.3)	**41.8** (+1.1)
		R@5 ↑	59.1	58.1 (-1.0)	**60.9** (+1.8)	67.9	67.8 (-0.1)	**68.4** (+0.5)
		R@10 ↑	70.5	69.5 (-1.0)	**72.0** (+1.5)	79.3	78.4 (-0.9)	**79.0** (-0.3)

Table D. **Detailed zero-shot retrieval performance on MSVD.**

Dataset	Pairs	Metic	Type	UMT-B			UMT-L		
				Base	ToMe	Ours	Base	ToMe	Ours
MSVD	5M	FLOPS ↓	(G)	78.5	66.2	50.6	267.8	182.1	165.3
		R@1 ↑	T2V	36.2	33.6	37.0	44.4	42.0	44.3
			V2T	58.5	57.5	59.1	66.1	65.2	68.2
		R@5 ↑	T2V	65.7	64.3	66.1	73.3	72.0	73.2
			V2T	78.7	77.3	78.8	85.5	85.8	86.0
		R@10 ↑	T2V	76.1	74.8	76.4	82.4	80.5	82.0
			V2T	84.3	84.8	84.3	89.4	89.6	90.2
	17M	FLOPS ↓	(G)	78.5	62.8	50.2	267.8	182.1	165.3
		R@1 ↑	T2V	41.4	38.6	42.2	49.9	48.1	50.0
			V2T	62.5	62.8	64.3	75.4	73.9	75.4
		R@5 ↑	T2V	70.6	69.0	71.3	77.7	76.6	77.6
			V2T	80.8	80.5	82.7	89.6	89.7	90.0
		R@10 ↑	T2V	80.1	78.9	80.5	85.3	84.3	85.5
			V2T	87.0	87.0	87.5	94.0	93.3	95.1
	25M	FLOPS ↓	(G)	78.5	66.2	50.6	267.8	182.1	153.6
		R@1 ↑	T2V	42.3	39.0	42.7	49.0	48.1	49.3
			V2T	61.9	61.2	64.5	74.5	72.4	75.7
		R@5 ↑	T2V	71.7	69.6	72.1	76.9	75.4	77.0
			V2T	82.5	82.5	84.0	89.7	89.6	89.4
		R@10 ↑	T2V	80.8	79.4	81.3	84.7	83.2	84.2
			V2T	88.5	88.8	89.4	92.8	93.3	93.3

Table E. **Detailed zero-shot retrieval performance on MSRVTT.**

Dataset	Pairs	Metic	Type	UMT-B			UMT-L		
				Base	ToMe	Ours	Base	ToMe	Ours
MSRVTT	5M	FLOPS ↓	(G)	78.5	66.2	50.2	267.8	182.1	164.8
		R@1 ↑	T2V	29.6	27.8	29.9	33.3	31.4	34.0
			V2T	26.2	25.0	25.4	30.2	29.4	31.3
		R@5 ↑	T2V	52.8	48.4	52.3	58.1	53.3	57.1
			V2T	46.7	44.9	47.3	51.3	51.4	53.0
		R@10 ↑	T2V	61.9	57.4	61.6	66.7	64.4	66.1
			V2T	54.9	54.3	55.0	61.6	61.4	63.1
	17M	FLOPS ↓	(G)	78.5	66.2	50.6	267.8	182.1	163.0
		R@1 ↑	T2V	35.5	32.5	36.5	42.6	39.7	42.1
			V2T	31.6	32.4	31.0	38.6	35.5	37.7
		R@5 ↑	T2V	59.3	55.8	59.6	64.4	62.0	63.9
			V2T	53.5	52.3	54.4	59.6	58.1	59.8
		R@10 ↑	T2V	68.6	64.1	68.9	73.1	70.5	72.4
			V2T	64.1	63.7	64.2	69.6	67.3	69.4
	25M	FLOPS ↓	(G)	78.5	66.2	44.7	267.8	160.9	146.5
		R@1 ↑	T2V	35.2	31.7	35.3	40.7	37.9	41.0
			V2T	30.3	30.4	31.3	37.1	34.6	37.4
		R@5 ↑	T2V	57.8	53.9	57.4	63.4	61.4	63.6
			V2T	50.7	51.7	52.0	58.7	56.6	58.8
		R@10 ↑	T2V	66.0	61.8	65.6	71.8	69.0	71.5
			V2T	61.4	62.2	62.8	68.9	66.2	68.6

Table F. **Detailed zero-shot retrieval performance on DiDeMo**

Dataset	#Pairs	Metic	Type	UMT-B			UMT-L		
				Base	ToMe	Ours	Base	ToMe	Ours
DiDeMo	5M	FLOPs ↓	(G)	78.5	66.2	47.3	267.8	182.1	166.2
		R@1 ↑	T2V	35.7	28.0	35.6	34.0	32.1	35.9
			V2T	35.8	31.4	36.1	36.2	35.3	37.2
		R@5 ↑	T2V	62.2	52.9	61.2	60.4	57.7	61.0
			V2T	61.6	58.1	60.0	60.0	62.3	63.2
		R@10 ↑	T2V	70.6	62.3	69.3	68.7	67.9	70.2
			V2T	69.6	67.9	68.7	68.6	71.2	71.9
	17M	FLOPs ↓	(G)	78.5	62.8	44.5	267.8	182.1	166.7
		R@1 ↑	T2V	41.9	36.9	42.6	46.4	44.8	49.7
			V2T	40.3	40.3	42.0	46.5	46.0	47.6
		R@5 ↑	T2V	66.7	63.0	68.6	70.0	72.3	73.9
			V2T	66.6	66.4	68.3	72.2	71.3	74.8
		R@10 ↑	T2V	75.0	71.9	77.2	78.8	80.2	80.9
			V2T	75.8	75.8	76.8	79.5	79.9	81.6
	25M	FLOPs ↓	(G)	78.5	62.8	48.2	267.8	160.9	162.5
		R@1 ↑	T2V	41.2	35.6	44.6	48.6	47.9	52.0
			V2T	40.8	41.5	43.4	49.9	48.4	52.0
		R@5 ↑	T2V	65.4	62.0	68.9	72.9	72.9	74.0
			V2T	67.7	67.0	68.9	74.8	74.1	75.9
		R@10 ↑	T2V	74.9	70.2	76.8	79.0	80.1	81.0
			V2T	76.7	75.9	78.6	81.4	81.6	83.8

Table G. **Detailed zero-shot retrieval performance on Activitynet.**

Dataset	#Pairs	Metic	Type	UMT-B			UMT-L		
				Base	ToMe	Ours	Base	ToMe	Ours
ANET	5M	FLOPs ↓	(G)	78.5	66.2	49.0	267.8	182.1	162.0
		R@1 ↑	T2V	28.3	24.7	29.3	31.9	29.4	31.4
			V2T	25.9	25.1	26.8	30.0	30.1	31.1
		R@5 ↑	T2V	53.0	48.9	53.2	60.2	57.4	59.6
			V2T	50.2	49.4	51.6	59.1	58.5	59.6
		R@10 ↑	T2V	64.2	60.5	64.3	72.0	69.6	71.5
			V2T	61.7	61.4	64.0	71.3	70.9	71.7
	17M	FLOPs ↓	(G)	78.5	62.8	45.0	267.8	182.1	151.7
		R@1 ↑	T2V	33.8	31.5	34.6	42.8	40.1	42.1
			V2T	31.6	32.0	33.2	40.7	40.4	41.2
		R@5 ↑	T2V	59.1	57.4	60.7	69.6	67.0	69.4
			V2T	56.2	57.5	58.5	67.6	67.2	68.2
		R@10 ↑	T2V	70.4	68.5	71.1	79.8	77.9	79.2
			V2T	67.9	69.2	70.3	78.6	78.5	79.1
	25M	FLOPs ↓	(G)	78.5	62.8	51.0	267.8	182.1	152.7
		R@1 ↑	T2V	35.5	32.6	36.4	41.9	41.6	42.8
			V2T	32.8	32.6	34.9	39.5	40.4	40.8
		R@5 ↑	T2V	60.6	57.8	61.3	68.6	68.1	68.8
			V2T	57.6	58.5	60.5	67.6	67.5	67.9
		R@10 ↑	T2V	71.8	69.0	72.4	79.6	78.7	79.6
			V2T	69.2	70.1	71.6	78.4	78.2	78.5

nual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011. 1

[5] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *ACL*, 2023. 1

[6] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *ICCV*, 2023. 1

[7] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *IJCV*, 2017. 1

[8] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1