



UNIVERSITÀ
DI TRENTO



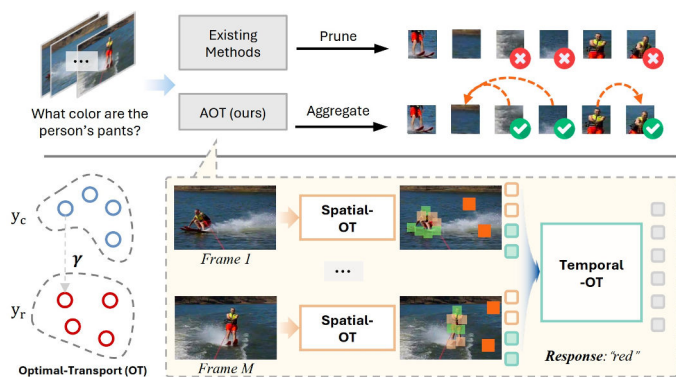
Token Reduction via Local and Global Contexts Optimization for Efficient Video Large Language Models

Jinlong Li, Liyuan Jiang, Haonan Zhang, Nicu Sebe



Introduction:

- Heavy redundant visual tokens among sampled video frames.
- Commonly focus intra-frame spatial redundancy, or prunes inside the LLM with shallow-layer overhead.
- Existing method simply removes or average up the similar visual tokens without in-depth exploitation.



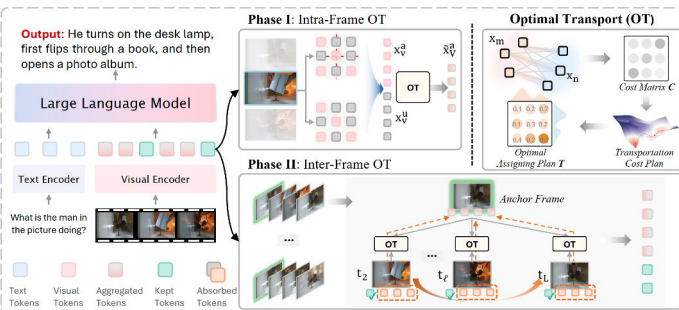
Optimal Transport

to deeply *exploit* and *aggregate* from the removed/averaged visual tokens

- ✓ **preserving temporal and visual fidelity**
- ✓ **Efficient and competitive**

Methodology

➤ Overview of AOT.



➤ Optimal Transport.

$$U = \sum_{m=1}^M u_m \delta_{x_m} \quad \text{and} \quad V = \sum_{n=1}^N v_n \delta_{x_n},$$

$$\langle T, C \rangle = \sum_{m=1}^M \sum_{n=1}^N T_{m,n} C_{m,n},$$

$$d_{OT,\lambda}(u, v|C) = \text{minimize } \langle T, C \rangle - \lambda h(T),$$

$$\text{subject to } T \mathbf{1}_N = u, T^\top \mathbf{1}_M = v, T \in \mathbb{R}_+^{M \times N}.$$

➤ Local-Global Token Anchors Establishment.

$$S_{[CLS]}^{\text{avg}} = \frac{1}{H} \sum_{h=1}^H S_{[CLS]}^h, \quad \mathbf{x}_V^1 = \bigcup_{w=1}^W \text{TopK}(\mathbf{x}_V^w, S_{[CLS]}^{\text{avg}}, K_w),$$

$$\mathbf{x}_V^g = \text{TopK}(\mathbf{x}_V, S_{[CLS]}^{\text{avg}}, K) \quad w=1$$

➤ Spatiotemporal Pruning.

$$d_{OT}^{\text{intra}}(k) = d_{OT}(u, v | \mathbf{1} - (\mathbf{X}_V^a)^\top (\mathbf{X}_V^u)),$$

$$d_{OT}^{\text{inter}}(\ell) = d_{OT}(u, v | \mathbf{1} - (\mathbf{A}^{(\ell-1)})(\mathbf{S}^{(\ell)})^\top),$$

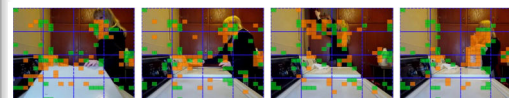
Experiments

Table 1. Comparison of state-of-the-art methods on LLaVA-OneVision [22] across video benchmarks. The best performance among those with similar retention ratios is highlighted in **bold**, while the second best will be denoted as underlined.

Method	Prefilling FLOPs (T) ↓	FLOPs Ratio ↓	Before LLM Retained Ratio	MVBench ↑	EgoSchema ↑	LongVideo Bench ↑	VideoMME ↑	Avg. ↑ Score	Avg. ↑ %
LLaVA-OV-7B	40.8	100%	100%	58.3	60.4	56.4	58.6	58.4	100
FastV [9]	9.3	22.8%	100%	55.9	57.5	56.7	56.1	56.5	96.7
PDrop [54]	10.5	25.7%	100%	56.1	58.0	54.1	56.4	56.2	96.2
DyCoke [44]	8.7	21.3%	25%	53.1	59.5	49.5	54.3	54.1	92.6
VisionZip [58]	8.7	21.3%	25%	<u>57.9</u>	<u>60.3</u>	<u>56.5</u>	58.2	<u>58.2</u>	<u>99.7</u>
PruneVid [17]	8.7	21.3%	25%	57.4	59.9	55.7	57.4	57.6	98.6
FastVID [40]	8.7	21.3%	25%	56.5	-	56.3	<u>58.0</u>	-	-
AOT	8.7	21.3%	25%	58.7	61.3	56.3	57.5	58.5	100.0
VisionZip [58]	7.0	17.2%	20%	57.2	59.8	55.2	57.9	57.7	98.8
PruneVid [17]	7.0	17.2%	20%	57.2	59.7	54.7	56.9	57.1	97.8
FastVID [40]	7.0	17.2%	20%	56.3	-	57.1	57.9	-	-
AOT	7.0	17.2%	20%	58.1	61.3	56.2	<u>57.2</u>	58.2	99.7
VisionZip [58]	5.2	12.7%	15%	56.5	59.8	54.4	56.1	56.7	97.1
PruneVid [17]	5.2	12.7%	15%	56.8	59.7	55.4	56.6	<u>57.1</u>	<u>97.8</u>
FastVID [40]	5.2	12.7%	15%	56.0	-	56.2	57.7	-	-
AOT	3.4	8.3%	15%	57.8	61.3	55.2	56.6	57.7	98.8
VisionZip [58]	3.4	8.3%	10%	53.5	58.0	49.3	53.4	53.5	91.6
PruneVid [17]	3.4	8.3%	10%	<u>56.2</u>	<u>59.8</u>	54.5	56.0	<u>56.6</u>	<u>96.9</u>
FastVID [40]	3.4	8.3%	10%	55.9	-	56.3	57.3	-	-
AOT	5.2	12.7%	10%	57.0	60.6	54.2	56.1	57.0	97.6

Table 2. Comparison of state-of-the-art methods on LLaVA-Video [70] across video benchmarks. The best performance among those highlighted in **bold**, while the second best will be denoted as underlined, demonstrating consistent effectiveness.

Method	Prefilling FLOPs (T) ↓	FLOPs Ratio ↓	Before LLM Retained Ratio	MVBench ↑	EgoSchema ↑	LongVideo Bench ↑	VideoMME ↑	Avg. ↑ Score	Avg. ↑ %
LLaVA-Video-7B	80.2	100%	100%	60.4	57.2	58.9	64.3	60.2	100
FastV [9]	17.1	21.3%	100%	54.3	54.1	55.0	58.8	55.6	92.4
PDrop [54]	19.5	24.3%	100%	55.9	54.3	54.7	61.9	56.7	94.2
VisionZip [58]	9.3	18.9%	25%	56.7	54.7	54.7	60.7	<u>56.7</u>	<u>94.2</u>
DyCoke [58]	9.3	18.9%	25%	50.8	-	53.0	56.9	-	-
AOT	9.3	18.9%	25%	58.8	55.4	56.2	62.4	58.2	96.7
VisionZip [58]	9.3	11.6%	15%	56.7	54.7	54.7	60.7	56.7	94.2
AOT	9.3	11.6%	15%	57.8	55.2	55.0	62.0	57.5	95.5



Q: What is the woman doing?
A: After closing the white box, the woman takes the food from cabinet and eats it.



Q: Is the boy talking with the man?
A: The boy wearing a white shirt is talking with the man.

