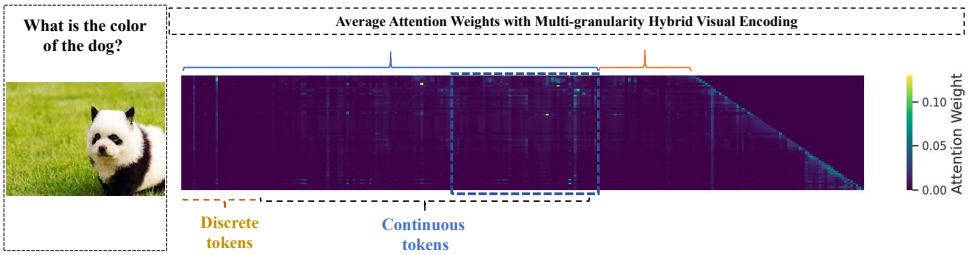

Figure 1: This image visualizes the attention map of MaVEn, which includes both visual discrete and continuous tokens, in the scenario where the user instruction inquires about fine grained detailed information in the image. We demonstrate the attention maps for the 31st layer.

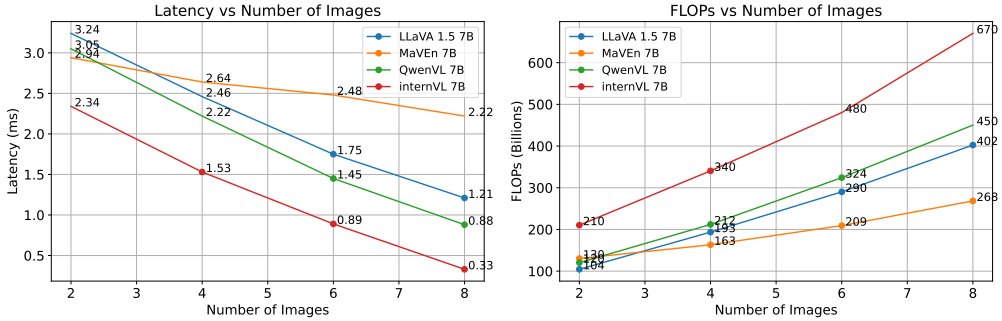

Figure 2: This figure visualizes the line charts of FLOPs and throughput varying with different numbers of image inputs during the inference stage for different MLLMs

Figure 3: This figure visualizes multiple images containing the discrete visual token 824, along with their visual discrete token sequences. We observe that all these images contain elements of the American flag.

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs
InternVL-Chat-V1.5	20B	**60.2**	**61.7**	**46.4**	**49.1**	**45.6**	**46.6**	**50.7**	**52.4**
Qwen-VL-Chat	7B	46.9	47.3	38.7	40.4	37.8	37.9	41.1	41.9
MaVEn	7B	50.3	51.4	37.2	38.4	35.3	36.5	40.9	42.1

Table 1: Performance of MLLMs on Video-MME

Stage	GPU	Time	Num of Image	Training Module	batch size
Stage1	8 × 80G A100	12h	1 million	Patch Selector	4096
Stage2	8 × 80G A100	84h	5 million	LLM Embedding Layer	128
Stage3	8 × 80G A100	8h	558K	Visual Projector	128
Stage3	8 × 80G A100	18h	707K	LLM & Visual Projector	128

Table 2: Training Cost for four-stage training