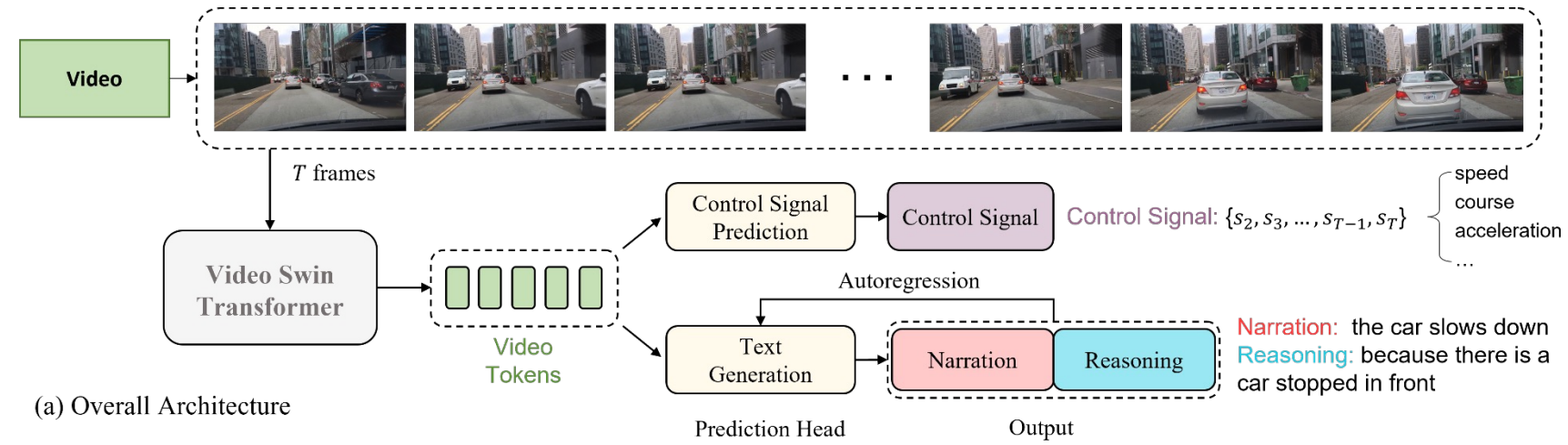
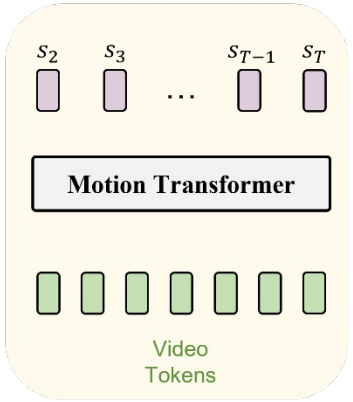
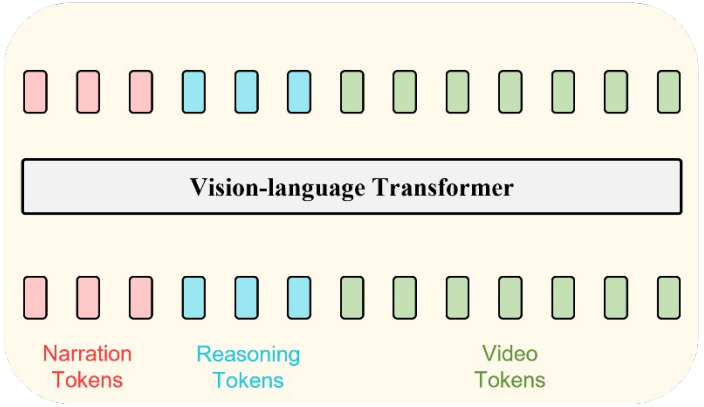
Implementation

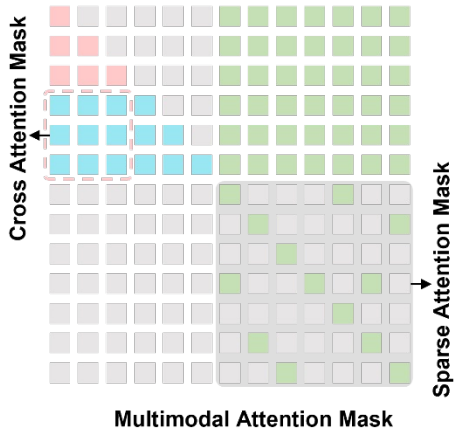


(a) Overall Architecture



(b) Motion Prediction Head





(c) Text Generation Head

Result

Method	Narration			Reasoning		
	B4	C	M	B4	C	M
S2VT [42]	30.2	179.8	27.5	6.3	53.4	11.2
S2VT++ [42]	27.1	157.0	26.4	5.8	52.7	10.9
SAA [56]	31.8	214.8	29.1	7.1	66.1	12.2
WAA [56]	32.3	215.8	29.2	7.3	69.5	12.2
Ours	**34.6**	**247.5**	**30.6**	**11.4**	**102.6**	**15.2**

Method	Narration	Reasoning	Full sentence
SAA [56]	90.8%	62.4%	-
WAA [56]	**93.5**%	66.0%	-
Ours	90.0%	**90.3**%	82.7%

Comparison with sota results: machine evaluation (top) and manual evaluation (bottom)

Ablation Experiment

Method	Narration				Reasoning			
	B4	C	M	R	B4	C	M	R
Single	33.2	238.9	29.7	62.0	8.6	89.7	14.1	31.4
Single+	33.9	**248.3**	30.5	**63.1**	9.3	97.2	14.6	31.5
Ours	**34.6**	247.5	**30.6**	62.8	**11.4**	**102.6**	**15.2**	**32.0**

Comparison of single-task and multi-task experiments

Ablation Experiment

Signals		Narration			Reasoning		
Speed	Course	C	M	R	C	M	R
✓		232.0	29.9	61.5	88.0	15.1	31.0
	✓	218.2	29.3	61.2	88.6	14.1	30.6
✓	✓	**247.5**	**30.6**	**62.8**	**102.6**	**15.2**	**32.0**

Effect of different control signals on caption results

Ablation Experiment

Method	Narration				Reasoning				Cost(min)
	B4	C	M	R	B4	C	M	R	
2	33.4	227.7	28.7	61.0	8.7	62.9	15.1	29.8	294
4	32.9	225.7	29.0	60.9	9.9	81.3	14.9	31.1	382
8	32.6	236.1	29.3	61.8	8.4	83.7	13.4	30.6	447
16	32.5	231.0	29.5	61.9	8.7	91.5	13.8	32.0	528
32	**34.6**	**247.5**	**30.6**	**62.8**	**11.4**	**102.6**	**15.2**	**32.0**	797

Effect of video frame number on the result

Ablation Experiment

$$c_\sigma = \begin{cases} 1, & -\sigma < \hat{c} - c < \sigma \\ 0, & otherwise \end{cases}$$

Method	Course						Speed					
	RMSE(degree)↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑	RMSE(m/s)↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑
Single	**6.3**	8.3	84.7	**90.5**	97.2	98.7	3.4	5.0	25.5	37.8	86.8	98.7
Ours	6.4	**62.2**	**85.5**	89.9	**97.2**	**98.8**	**2.5**	**11.1**	**28.1**	**45.3**	**94.3**	**99.5**

Effect of Multitasking on Control Signal Prediction

Visualization