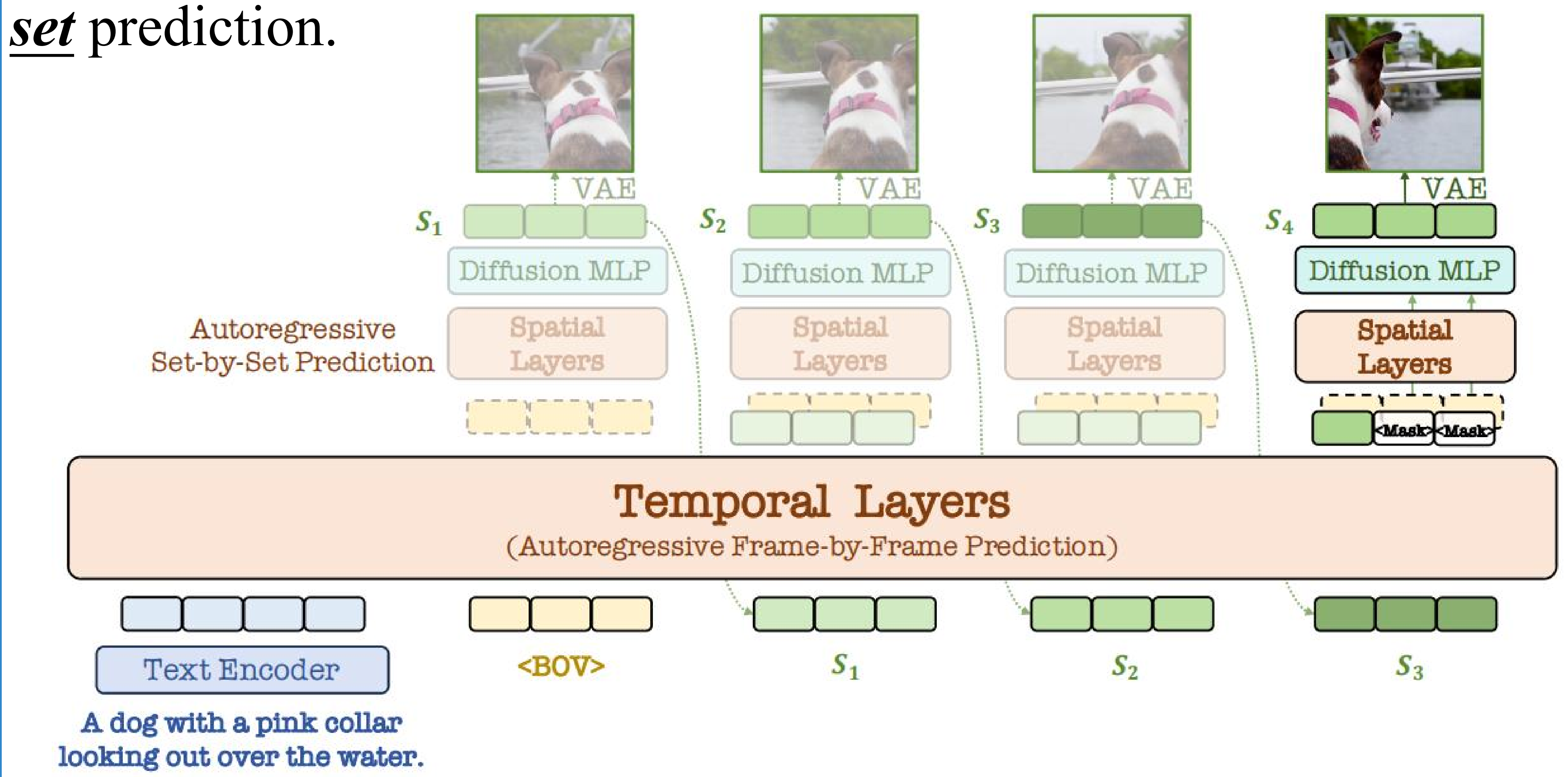
Autoregressive Video Generation without Vector Quantization

Haoge Deng[1,5*], Ting Pan[2,3,5*], Haiwen Diao[3,5*], Zhengxiong Luo[5*], Yufeng Cui[5],
Huchuan Lu[4], Shiguang Shan[2,3], Yonggang Qi[1✉], Xinlong Wang[5✉]

Wechat Github

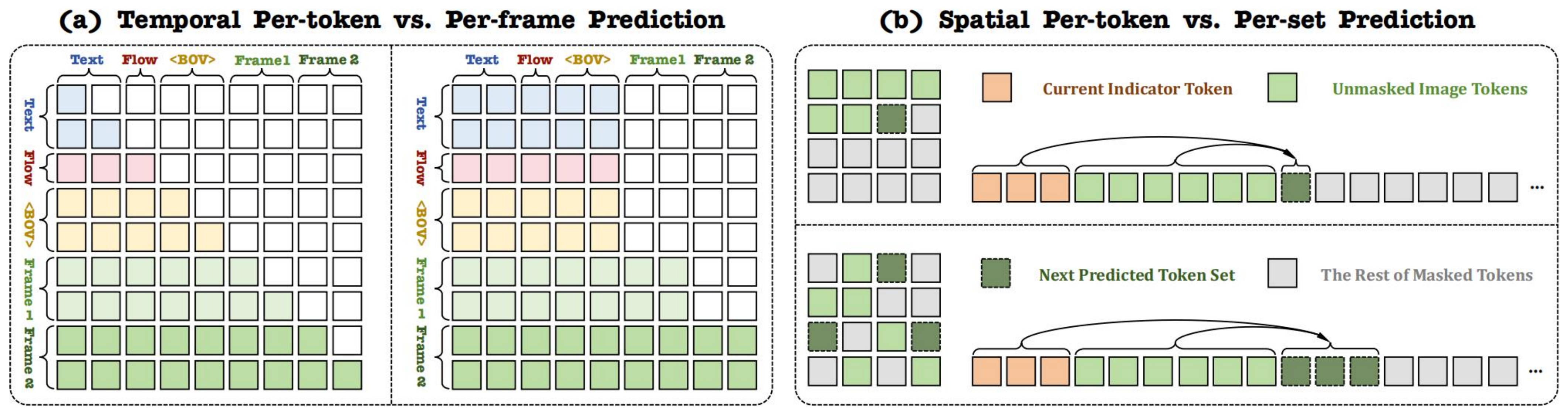
Introduction

Unlike raster-scan prediction in prior AR model or joint distribution modeling of fixed-length tokens in DM. We propose to reformulate the video generation problem as a non-quantized autoregressive modeling of ***temporal frame-by-frame*** prediction and ***spatial set-by-set*** prediction.



Attention Mechanism

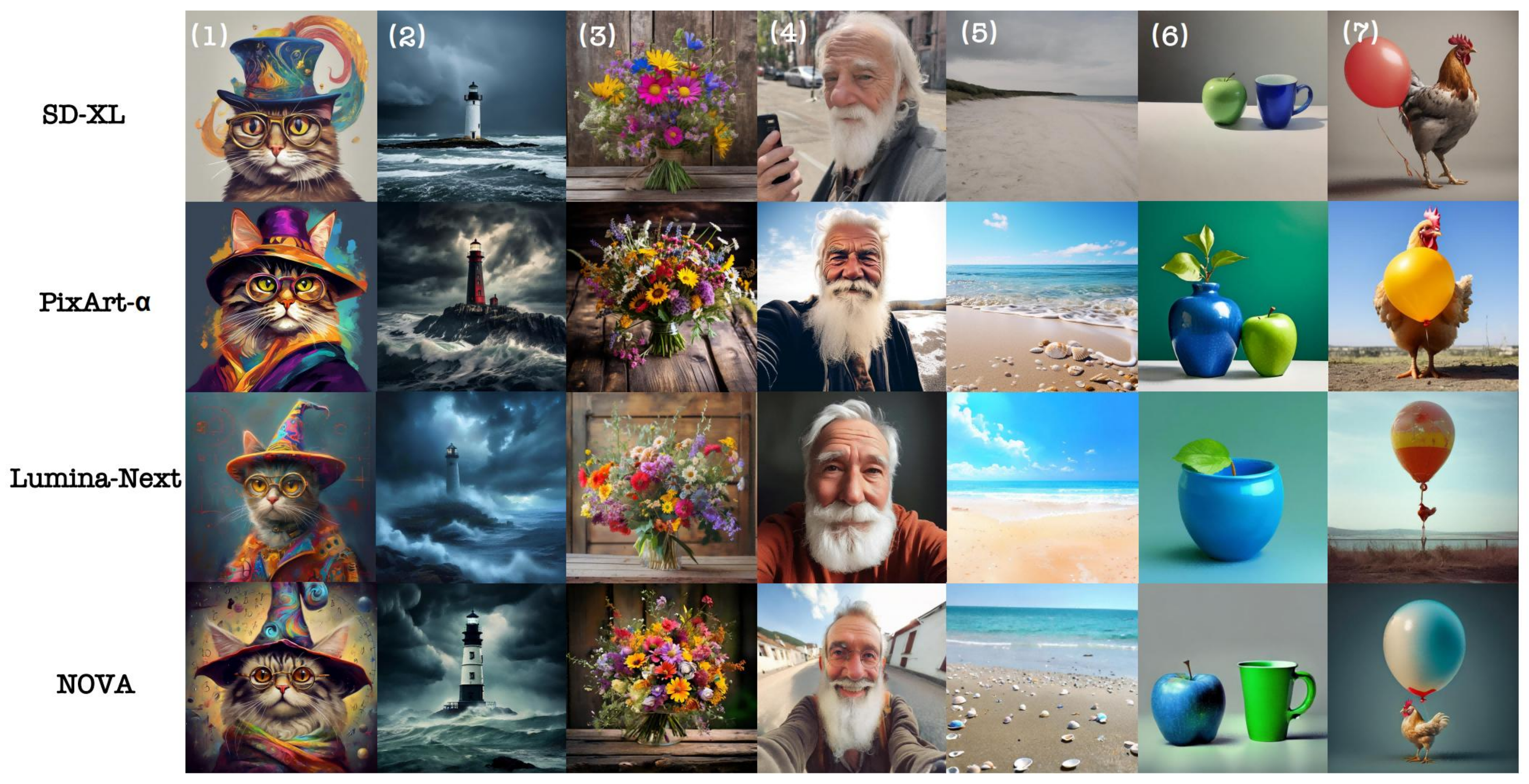
➢NOVA regressively predicts each frame in ***a casual order across the temporal scale***, and predicts each token set in ***a random order across the spatial scale***.



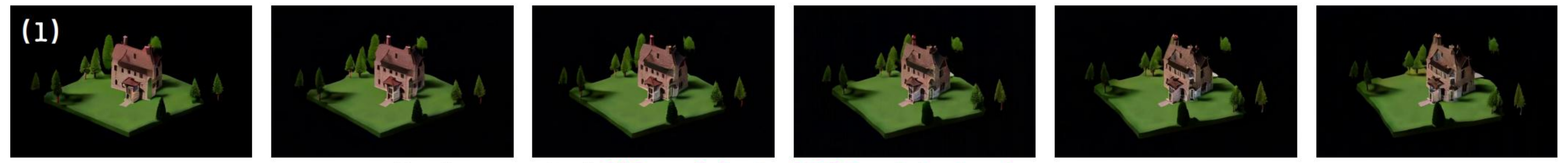
Results

Qualitative Results

➢Text-To-Image



➢Text-To-Video



Text prompt : A 3D model of a 1800s victorian house.

Text prompt : A cat wearing sunglasses and working as a lifeguard at a pool.

Text prompt : A drone view of celebration with Christmas tree and fireworks, starry sky, background.

* Equal contribution
1 Beijing University of Posts and Telecommunications
2 Key Laboratory of Intelligent Information Processing, ICT, CAS
3 University of Chinese Academy of Sciences
4 Dalian University of Technology 5 Beijing Academy of Artificial Intelligence

Quantitative Results

➢ NOVA outperforms existing text-to-image models with superior performance and efficiency.

Model	ModelSpec		T2I-CompBench			GenEval						DPG-Bench	A100 days
	#params	#images	Color	Shape	Texture	Overall	Single	Two	Counting	Colors	Position	ColorAttr	Overall	
Diffusion models													
PixArt-α	0.6B	25M	68.86	55.82	70.44	0.48	0.98	0.50	0.44	0.80	0.08	0.07	71.11	753
SD v1.5	1B	2B	37.50	37.24	42.19	0.43	0.97	0.38	0.35	0.76	0.04	0.06	63.18	-
SD v2.1	1B	2B	56.94	44.95	49.82	0.50	0.98	0.37	0.44	0.85	0.07	0.17	-	-
SDXL	2.6B	-	63.69	54.08	56.37	0.55	0.98	0.44	0.39	0.85	0.15	0.23	74.65	-
DALL-E2	6.5B	650M	57.50	54.64	63.74	0.52	0.94	0.66	0.49	0.77	0.10	0.19	-	-
DALL-E3	-	-	81.10	67.50	80.70	0.67	0.96	0.87	0.47	0.83	0.43	0.45	83.50	-
SD3	2B	-	-	-	-	0.62	0.98	0.74	0.63	0.67	0.34	0.36	84.10	-
Autoregressive models													
LlamaGen	0.8B	60M	-	-	-	0.32	0.71	0.34	0.21	0.58	0.07	0.04	-	-
Emu3 (+ Rewriter)	8B	-	79.13	58.46	74.22	0.66	0.99	0.81	0.42	0.80	0.49	0.45	81.60	-
NOVA (512×512)	0.6B	16M	70.75	55.98	69.79	0.66	0.98	0.85	0.58	0.83	0.20	0.48	81.76	127
+ Rewriter	0.6B	16M	83.02	61.47	75.80	0.75	0.98	0.88	0.62	0.82	0.62	0.58	-	127
+ Videos	0.6B	36M	71.80	47.86	65.31	0.55	0.98	0.56	0.48	0.75	0.15	0.41	81.77	342
+ Videos & Rewriter	0.6B	36M	81.36	59.16	72.45	0.71	0.98	0.83	0.52	0.81	0.58	0.51	-	342
NOVA (1024×1024)	0.3B	600M	73.35	57.28	70.09	0.67	0.98	0.86	0.53	0.84	0.32	0.52	80.60	267
NOVA (1024×1024)	0.6B	600M	74.72	56.99	69.50	0.69	0.98	0.89	0.56	0.84	0.32	0.56	82.25	320
NOVA (1024×1024)	1.4B	600M	74.30	57.14	70.00	0.71	0.99	0.91	0.62	0.85	0.33	0.56	83.01	608

➢NOVA rivals diffusion text-to-video models and significantly suppresses the AR counterpart.

Model	#params	#videos	latency	Total Score	Quality Score	Semantic Score	Aesthetic Quality	Object Class	Multiple Objects	Human Action	Spatial Relationship	Scene
Closed-source models												
Gen-2	-	-	-	80.58	82.47	73.03	66.96	90.92	55.47	89.20	66.91	48.91
Kling (2024-07)	-	-	-	81.85	83.39	75.68	61.21	87.24	68.05	93.40	73.03	50.86
Gen-3	-	-	-	82.32	84.11	75.17	63.34	87.81	53.64	96.4	65.09	54.57
Diffusion models (w/ SD init)												
LaVie	3B	25M	-	77.08	78.78	70.31	54.94	91.82	33.32	96.8	34.09	52.69
Show-1	4B	10M	-	78.93	80.42	72.98	57.35	93.07	45.47	95.60	53.50	47.03
AnimateDiff-v2	1B	10M	-	80.27	82.90	69.75	67.16	90.90	36.88	92.60	34.60	50.19
VideoCrafter-v2.0	2B	10M	-	80.44	82.20	73.42	63.13	92.55	40.66	95.00	35.86	55.29
T2V-Turbo (VC2)	2B	10M	-	81.01	82.57	74.76	63.04	93.96	54.65	95.20	38.67	55.58
Diffusion models												
OpenSora-v1.1	1B	10M	48s	75.66	77.74	67.36	50.12	86.76	40.97	84.20	52.47	38.63
OpenSoraPlan-v1.1	1B	4.5M	60s	78.00	80.91	66.38	56.85	76.30	40.35	86.80	53.11	27.17
OpenSora-v1.2	1B	32M	55s	79.76	81.35	73.39	56.85	82.22	51.83	91.20	68.56	42.44
CogVideoX	2B	35M	90s	80.91	82.18	75.83	60.82	83.37	62.63	98.00	69.90	51.14
Autoregressive models												
CogVideo	9B	5.4M	-	67.01	72.06	46.83	38.18	73.4	18.11	78.20	18.24	28.24
Emu3	8B	-	-	80.96	84.09	68.43	59.64	86.17	44.64	77.71	68.73	37.11
NOVA	0.6B	20M	12s	78.48	78.96	76.57	54.52	91.36	73.46	91.20	66.37	50.16
+ Rewriter	0.6B	20M	12s	80.12	80.39	79.05	59.42	92.00	77.52	95.20	77.52	54.06