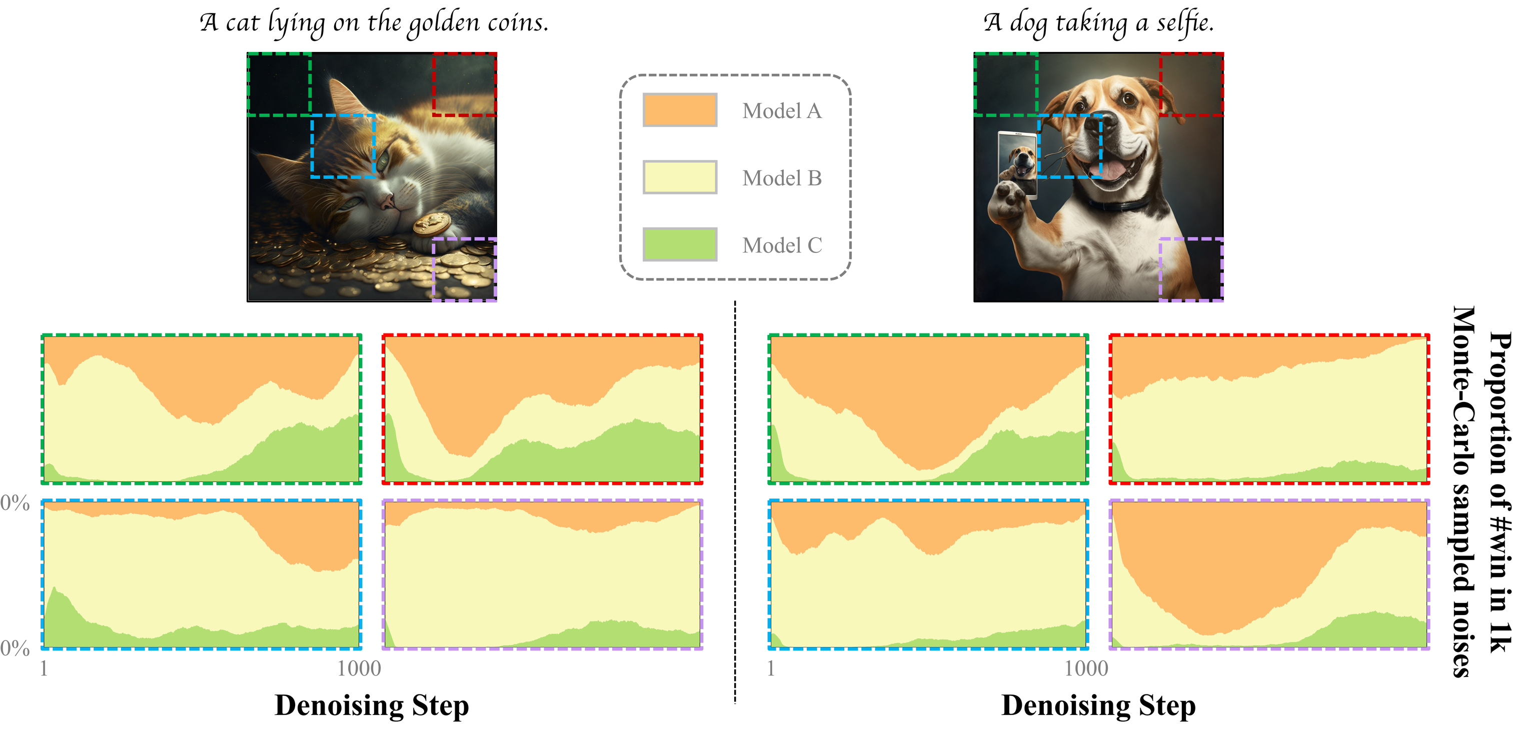
Ensembling Diffusion Models via Adaptive Feature Aggregation

Cong Wang[1*], Kuan Tian[2*], Yonghang Guan[2], Fei Shen[2], Zhiwei Jiang[1†], Qing Gu[1], Jun Zhang[2†]

[1] Nanjing University, [2] Tencent AIPD; * Equal Contribution, † Corresponding Author

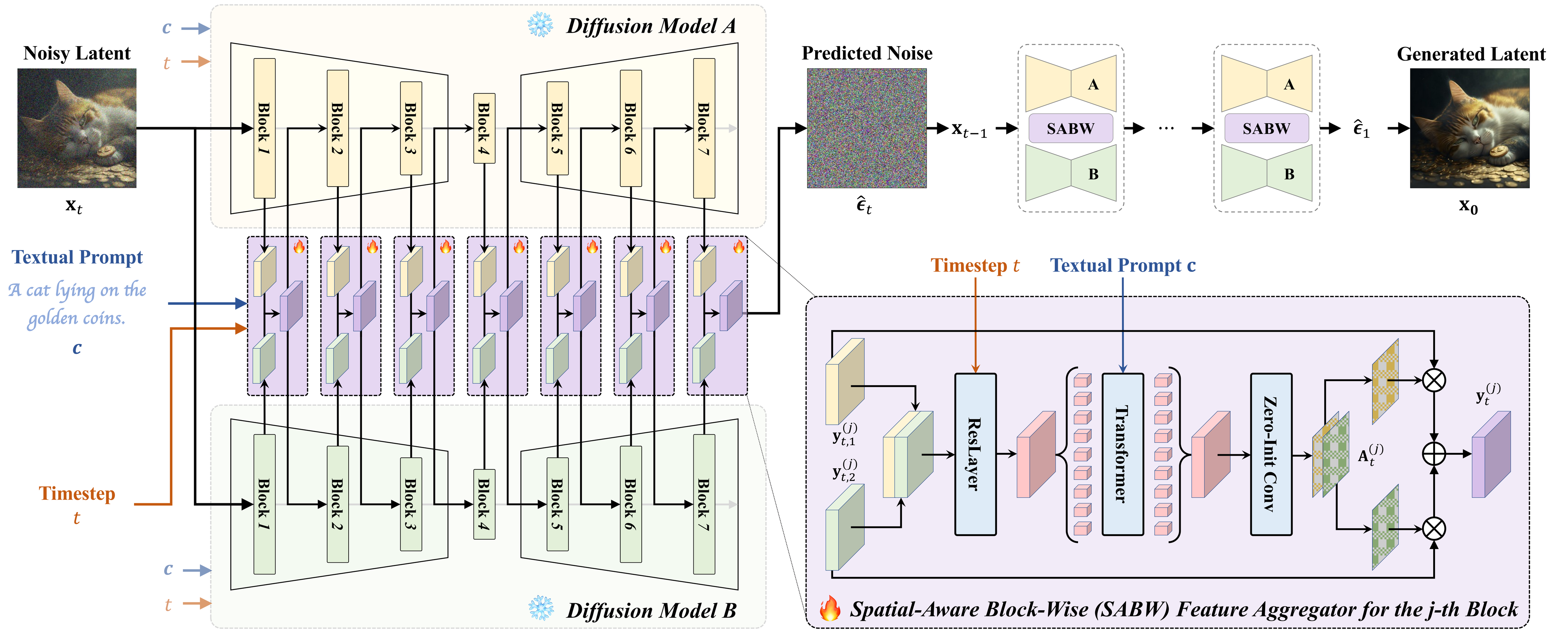
I. INTRODUCTION

- We aims to leverage **multiple** text2image **diffusion** models to dig out better generation.
- The **prompts**, **noises**, **timesteps**, and **spatial locations** have an impact on the denoising capabilities.
- We propose **Adaptive Feature Aggregation (AFA),** which dynamically adjusts contributions of multiple models at **feature level** by taking into account various states.



II. METHOD

- **AFA** ensembles multiple diffusion models that share same architecture but different parameters.
- **SABW** feature aggregator is learned to aggregate output features of each block from multiple U-Net denoisers.
- Only **SABW feature aggregators** are **trained**, while **denoisers** are **frozen**.



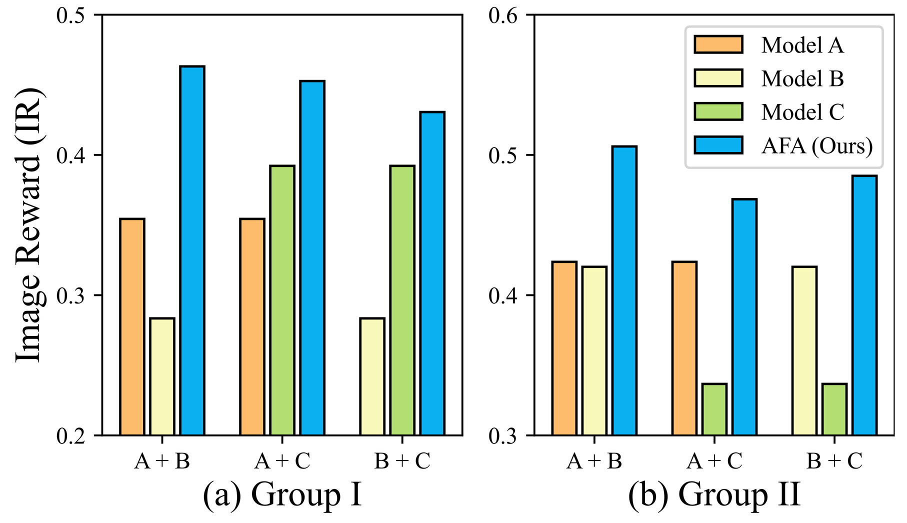
III. EXPERIMENTS

	COCO 2017				Draw Bench Prompts			
	FID ↓	IS	CLIP-I	CLIP-T	AES	PS	HPSv2	IR
Base Model A	13.01	5.65	.6724	.2609	5.4102	21.6279	27.8007	.3544
Base Model B	13.45	5.43	.6775	.2652	5.5013	21.4624	27.7246	.2835
Base Model C	12.32	6.32	.6890	.2566	5.4881	21.8031	27.9652	.3922
Wtd. Merging	10.65	6.93	.6861	.2626	5.4815	21.7272	27.9086	.3909
MBW	11.03	6.51	.6870	.2624	5.4812	21.7201	27.9080	.3922
autoMBW	13.35	5.51	.6772	.2577	5.5056	21.4785	27.8192	.3672
MagicFusion	10.53	6.85	.6751	.2620	5.3431	21.3840	27.8105	.3317
AFA (Ours)	**9.76**	**7.14**	**.6926**	**.2675**	**5.5201**	**21.8263**	**27.9734**	**.4388**

Quantitative comparison for ***ER***, ***MMR***, and ***RV***.

	COCO 2017				Draw Bench Prompts			
	FID ↓	IS	CLIP-I	CLIP-T	AES	PS	HPSv2	IR
Base Model A	12.12	5.66	.6849	.2623	5.5641	21.8013	28.0183	.4238
Base Model B	12.41	5.59	.6818	.2580	5.5027	21.7249	28.0343	.4202
Base Model C	12.05	5.95	.6638	.2642	5.5712	21.4936	27.8089	.3367
Wtd. Merging	11.53	6.56	.6824	.2631	5.5756	21.7516	28.0014	.4387
MBW	12.06	6.42	.6826	.2632	5.5772	21.7487	28.0029	.4396
autoMBW	12.39	5.62	.6774	.2588	5.5478	21.5135	27.9873	.3513
MagicFusion	11.63	7.13	.6790	.2640	5.4674	21.4270	27.9608	.4194
AFA (Ours)	**10.27**	**7.42**	**.6855**	**.2717**	**5.5798**	**21.8059**	**28.0371**	**.4892**

Quantitative comparison for ***AR***, ***CR***, and ***RCR***.

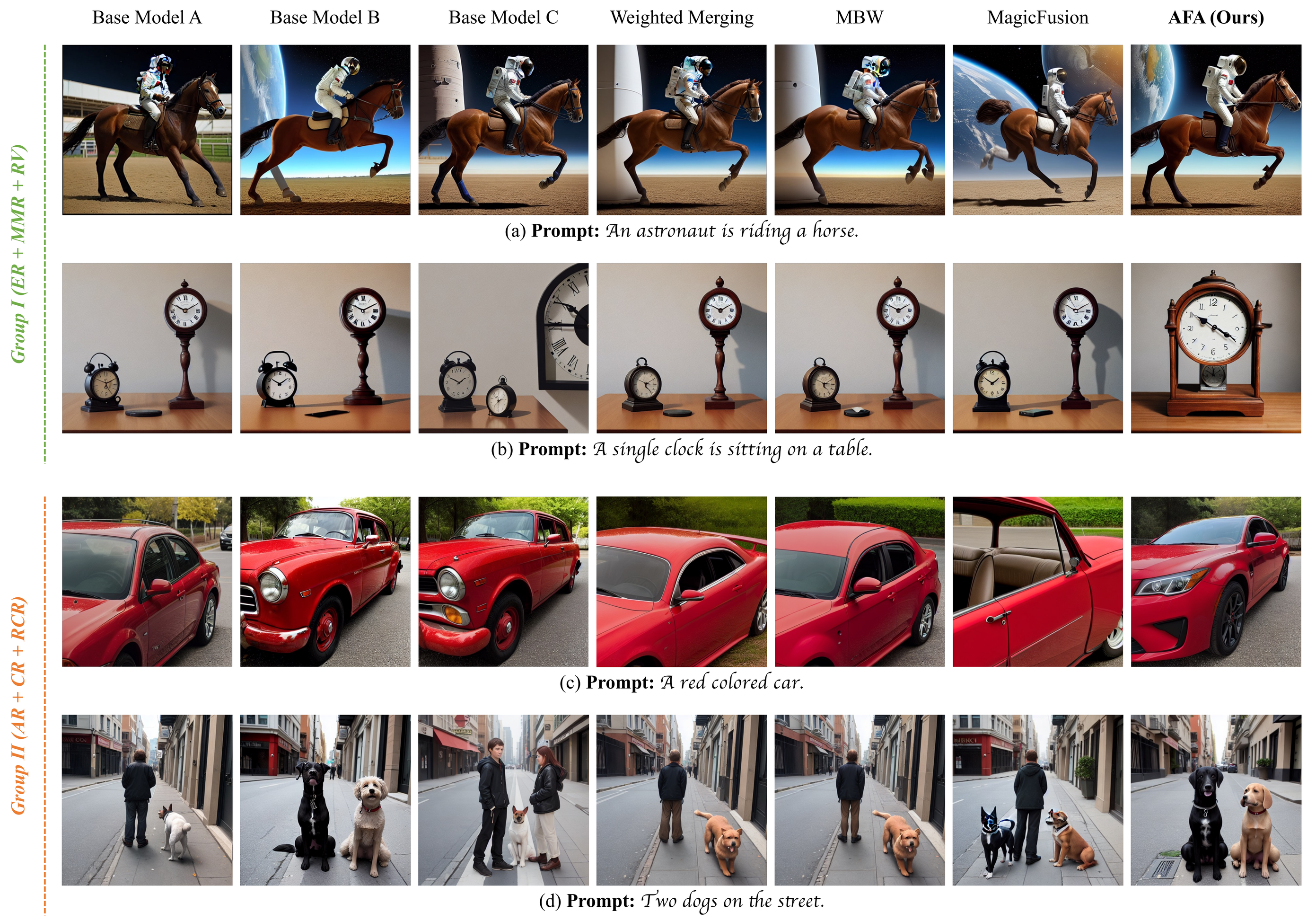


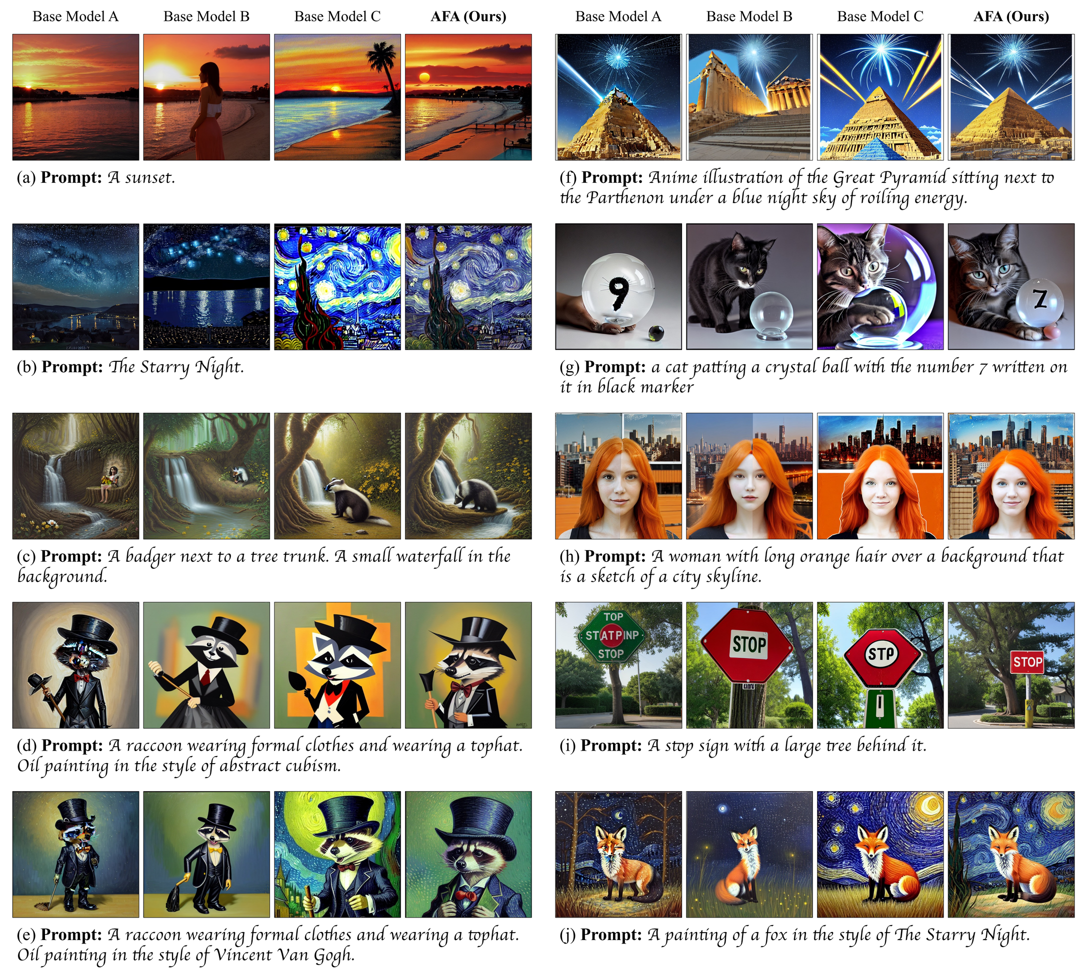
Quantitative comparison with ***two base models***.

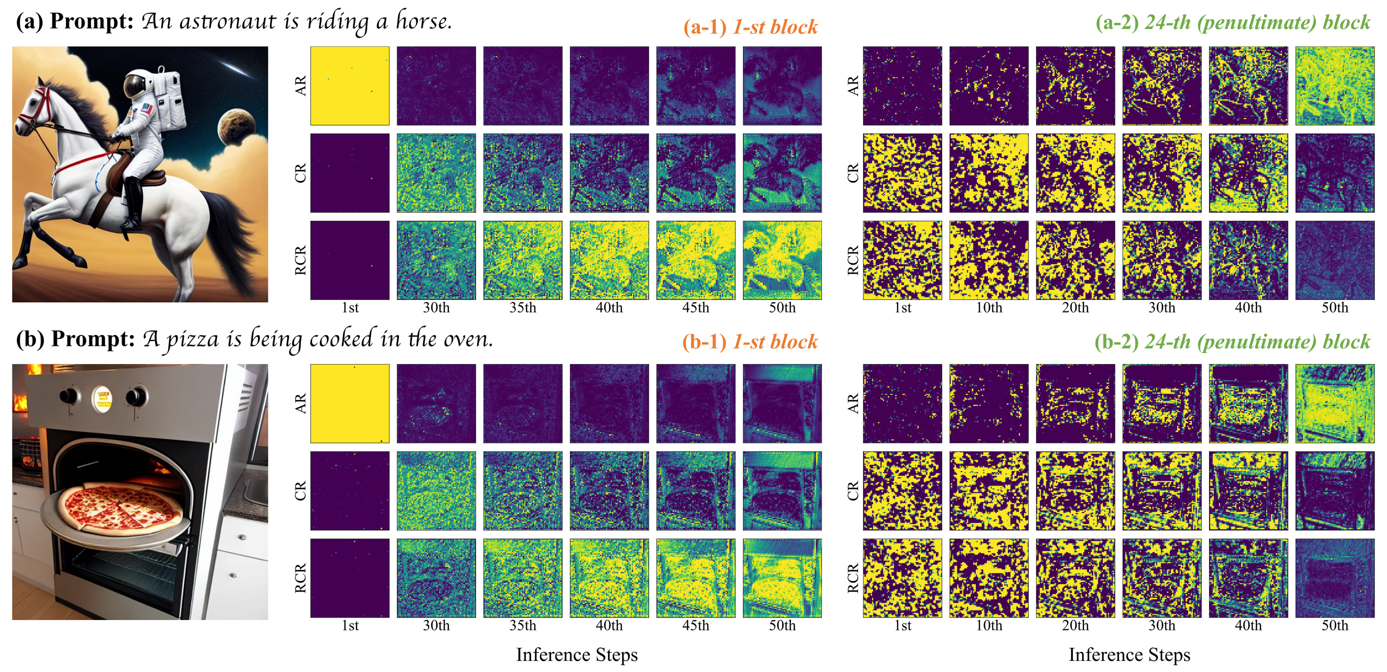
	IR (Group I)	IR (Group II)
AFA (Full Model)	**.4388**	**.4892**
Only Ensembling Last Block	.4176	.4374
Block-Wise Averaging	.4001	.4372
Noise Averaging	.3919	.4355
w/o Spatial Location	.4044	.4610
w/o Timestep	.4235	.4622
w/o Textual Condition	.4163	.4559

Ablation study of AFA.

IV. VISUALIZATION & ANALYSIS

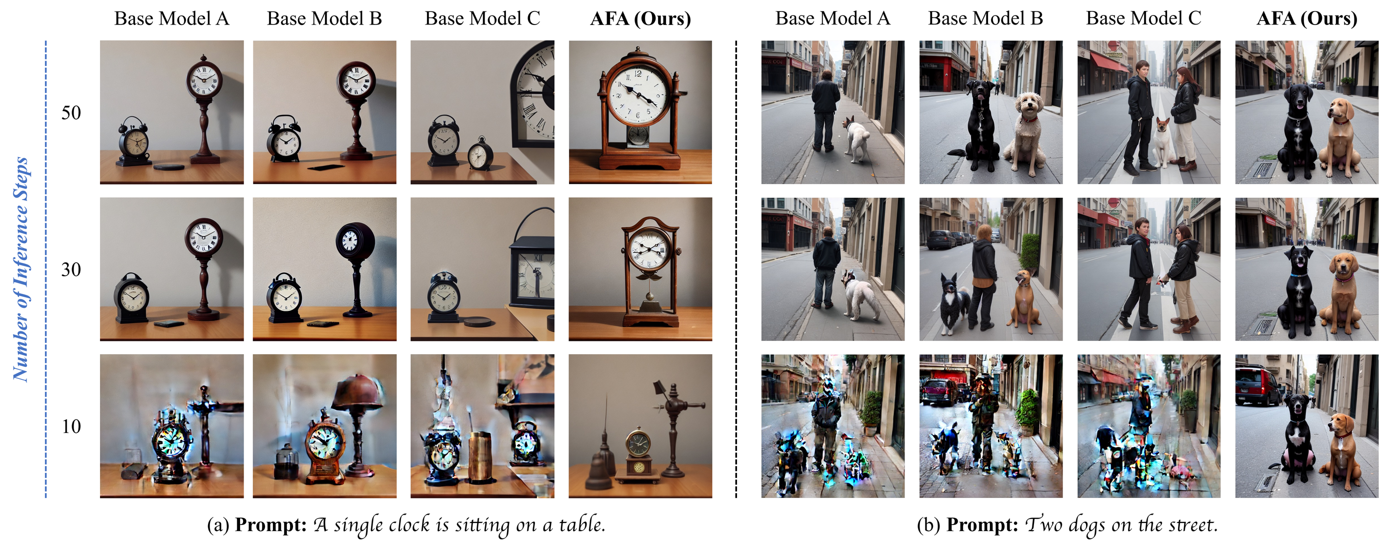


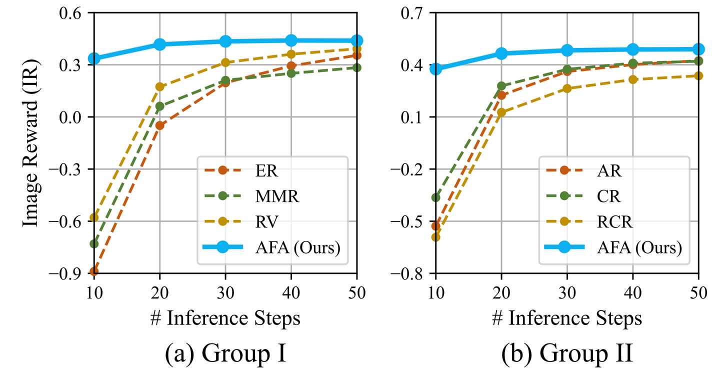




Visualization of the learned attention maps.

This indicates that AFA can effectively ensemble diffusion models based on context and timesteps.





Greater tolerance for fewer inference steps.