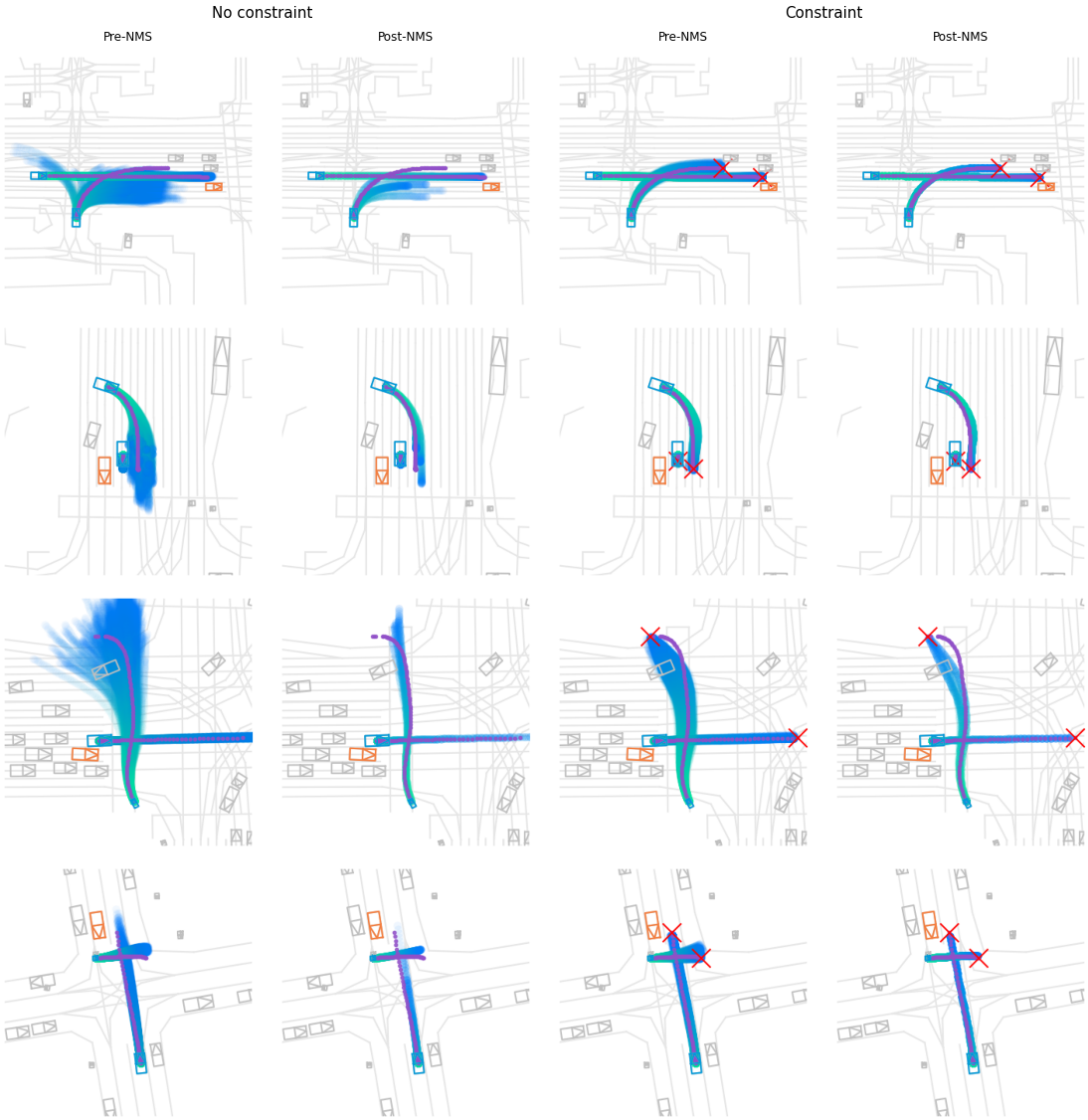
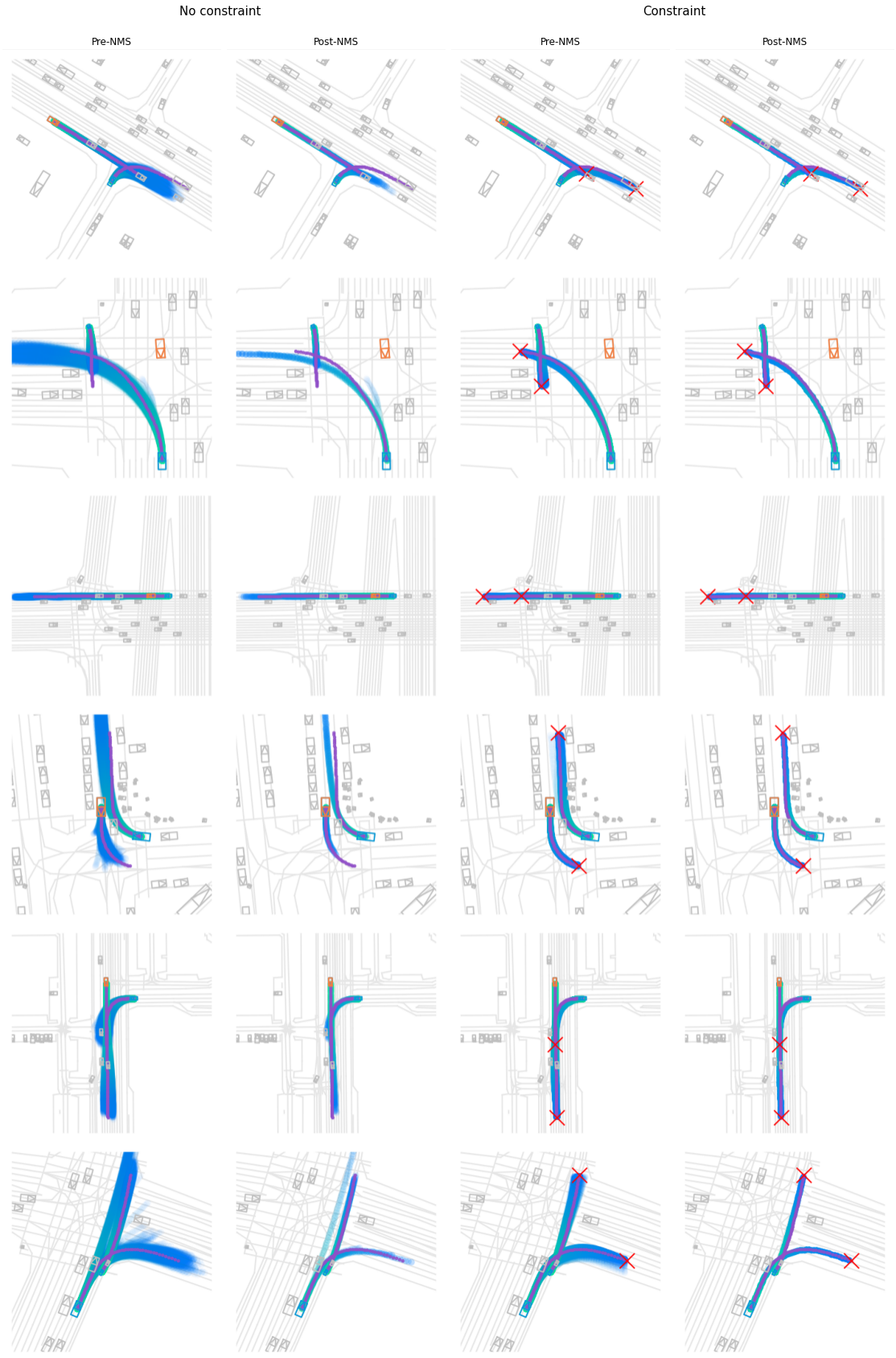
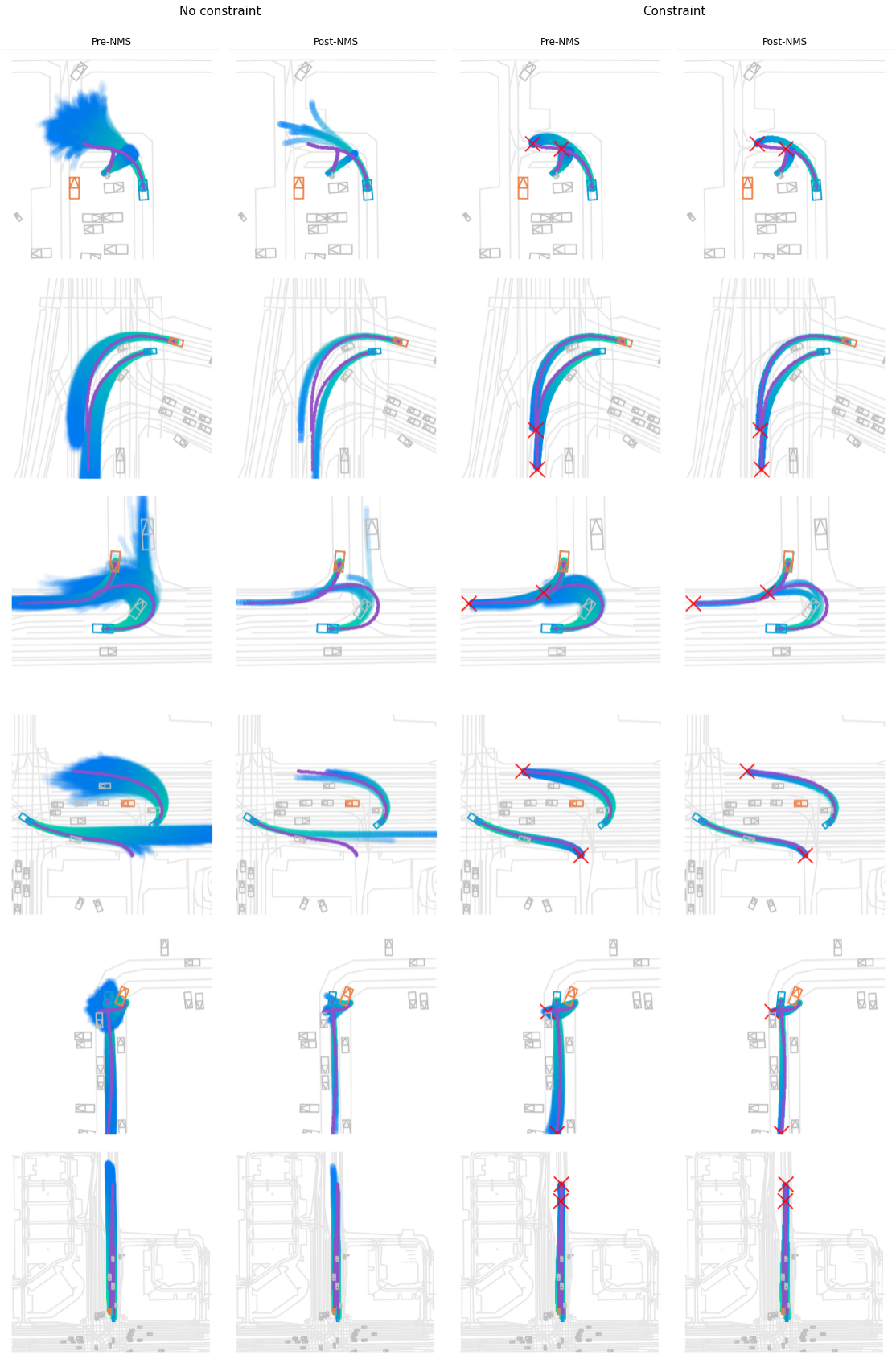
MotionDiffuser: Controllable Multi-Agent Motion Prediction using Diffusion

Chiyu "Max" Jiang* Andre Cornman* Cheolho Park
Ben Sapp Yin Zhou Dragomir Anguelov
* equal contribution
Waymo LLC

1. Additional Visualizations




No constraint
Constraint
Pre-NMS
Post-NMS
Pre-NMS
Post-NMS


No constraint
Constraint
Pre-NMS
Post-NMS
Pre-NMS
Post-NMS

2. Implementation Details

MotionDiffuser is trained on the Waymo Open Motion Dataset using 32 TPU shards for $2 * 10^6$ training steps. We use the ADAMW optimizer [2] with weight decay coefficient of 0.03. The learning rate is set to $5 * 10^{-4}$, with 10^4 warmup steps and linear learning rate decay. MotionDiffuser uses the Wayformer [3] encoder backbone, with 128 latent embeddings, each with hidden size of 256. Because the Wayformer encoder is agent centric, we append each agent's position and heading (relative to the ego vehicle) to its corresponding context vectors.

Our transformer denoiser architecture uses 4 layers of self-attention and cross-attention blocks. Each attention layer has a hidden size of 256 and an intermediate size of 1024. ReLU activation is used in all transformer layers. We embed the noise level using 128 random fourier features.

We can flexibly denoise N random noise vectors during training and inference. We use $N = 128$ during training and $N = 256$ during inference (before applying clustering).

3. Network Preconditioning

We follow the network preconditioning framework from [1], which defines the denoiser $D_{\boldsymbol{\theta}}$ as:

$$D_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{c}, \sigma) = c_{\text{skip}}(\sigma)\boldsymbol{x} + c_{\text{out}}(\sigma)F_{\boldsymbol{\theta}}(c_{\text{in}}(\sigma)\boldsymbol{x}; \boldsymbol{c}, c_{\text{noise}}(\sigma)) \tag{1}$$

$c_{\text{in}}(\sigma)$ scales the network input, such that the training inputs to $F_{\boldsymbol{\theta}}$ have unit variance.

$$c_{\text{in}}(\sigma) = 1/\sqrt{\sigma^2 + \sigma_{data}^2} \tag{2}$$

$c_{\text{skip}}(\sigma)$ modulates the skip connection and is defined as:

$$c_{\text{skip}}(\sigma) = \sigma_{data}^2/(\sigma^2 + \sigma_{data}^2) \tag{3}$$

$c_{\text{out}}(\sigma)$ modulates the network output and is defined as:

$$c_{\text{out}}(\sigma) = \sigma \cdot \sigma_{data}/\sqrt{\sigma^2 + \sigma_{data}^2} \tag{4}$$

Finally $c_{\text{noise}}(\sigma)$ scales the noise level, and is defined as:

$$c_{\text{noise}}(\sigma) = \frac{1}{4}\ln\sigma \tag{5}$$

For all our experiments, we set $\sigma_{data} = 0.5$.

4. Inference Latency

We report our model's inference latency over a varying number of sampling steps T in Table 1. We use a single V100 GPU, with batch size of 1.

Method	Latency (ms)	minSADE(↓)	minSFDE(↓)	SMissRate(↓)
Ours ($T = 8$)	101.0	0.91	2.06	0.47
Ours ($T = 16$)	203.7	0.88	1.96	0.44
Ours ($T = 32$)	408.5	0.88	1.97	0.43

Table 1. Model inference latency vs. quality for WOMD Interactive Validation Split.

References

[1] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 4

[2] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 4

[3] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. *arXiv preprint arXiv:2207.05844*, 2022. 4