

Supplementary Material: Exploring Masked Diffusion Transformers for Co-Speech Gesture Generation

A Derivation of THEOREM 4.1

The proof is based on [2] and [3].

The sample \hat{x}_0 predicted by the network during the sampling phase and the sample x_0 from the training phase is not identical; that is, $|\hat{x}_0 - x_0|_1$ is non-zero. This discrepancy is more apparent when the number of sampling iterations is reduced. However, $p_\theta(x_{t-1}|x_t) = p(x_{t-1}|x_t, x_0)$ holds true when $\hat{x}_0 = x_0$, this requires the network to have no prediction error for x_0 . In order to minimize errors during the sampling process, especially under conditions of fewer network predictions, we have conducted some derivations.

To start, we have the following definition: according to analysis-dpm [1], we model \hat{x}_0 as $\hat{x}_0 = x_0 + e_t * \epsilon_0$, $\epsilon_0 \sim \mathcal{N}(0, I)$. For $p_\theta(x_{t-1}|x_t)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (1)$$

Where $\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\hat{x}_0 + \frac{\sqrt{\alpha_t}(1-\alpha_{t-1})}{1-\alpha_t}x_t$. We consider how much error is introduced by the inconsistency between \hat{x}_0 and x_0 . We substitute $\hat{x}_0 = x_0 + e_t * \epsilon_0$ into Eq. 1. Since the mean of ϵ_0 is zero, the mean of $p_\theta(x_{t-1}|x_t)$ remains unchanged. We only consider the variance of $p_\theta(x_{t-1}|x_t)$:

$$\text{Var}(\hat{x}_t) = \text{Var}\left(\frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\hat{x}_0\right) + \Sigma_\theta(x_t, t) \quad (2)$$

$$= \text{Var}\left(\frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}(x_0 + e_t * \epsilon_0)\right) + \Sigma(x_t, t) \quad (3)$$

$$= \text{Var}\left(\frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}(x_0 + e_t * \epsilon_0)\right) + \Sigma(x_t, t) \quad (4)$$

$$= \left(\frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}e_t\right)^2 + \Sigma(x_t, t) \quad (5)$$

During inference, we utilize a fixed variance, resulting in $\Sigma_\theta(x_t, t) = \Sigma(x_t, t)$. Table 1 shows the discrepancy between $p_\theta(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t, x_0)$.

Table 1: The distribution $p(x_{t-1}|x_t, x_0)$ during training and $p_\theta(x_{t-1}|x_t)$ during DDPM sampling.

	Mean	Variance
$p(x_{t-1} x_t, x_0)$	$\mu_\theta(x_t, t)$	$(\Sigma(x_t, t))I$
$p_\theta(x_{t-1} x_t)$	$\mu_\theta(x_t, t)$	$(\Sigma(x_t, t) + (\frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}e_t)^2)I$

From Table 1, it can be seen that the variance of the sampling distribution $p_\theta(x_{t-1}|x_t)$ is always greater than the variance of the training distribution $p(x_{t-1}|x_t, x_0)$. [3] pointed out that this error cannot be addressed by adjusting the variance of $p_\theta(x_{t-1}|x_t)$, but can be mitigated by scaling the predicted noise ϵ_t in the network, this indirectly reduces the variance of $p_\theta(x_{t-1}|x_t)$. However, since our MDT-A2G network predicts the original signal \hat{x}_0 , assume that \hat{x}_0^t represents the original gesture predicted by the diffusion model

Table 2: Quantitative comparison with different accelerating configurations.

Method	DDIM 50 steps	Acceleration ratio	scale	Average Time(s)↓	FGD↓	Diversity→
MDT-A2G-B	✓	×	×	0.516 ± 0.099	161.81	137.33
MDT-A2G-B	×	1:20	1	1.984 ± 0.214	59.93	334.88
MDT-A2G-B	×	1:20	1.0005	1.984 ± 0.215	57.61	340.11

at step t , taking into account that $\epsilon_t = (\hat{x}_t - \sqrt{\alpha_t}\hat{x}_0^t)/\sqrt{1-\alpha_t}$, some transformations are needed:

$$\epsilon_t = (\hat{x}_t - \sqrt{\alpha_t}\hat{x}_0^t)/\sqrt{1-\alpha_t} \quad (6)$$

$$\hat{x}_0^{t-1} = (\hat{x}_t - \sqrt{1-\alpha_t} * \frac{\epsilon_t}{scale})/\sqrt{\alpha_t} \quad (7)$$

$$\hat{x}_0^{t-1} = (\hat{x}_t - \sqrt{1-\alpha_t} * \frac{(\hat{x}_t - \sqrt{\alpha_t}\hat{x}_0^t)/\sqrt{1-\alpha_t}}{scale})/\sqrt{\alpha_t} \quad (8)$$

Ensuring that *scale* is greater than 1 reduces the variance of $p_\theta(x_{t-1}|x_t)$. Based on experience, we set *scale* to 1.0005. In the acceleration phase, we scale the variance of the previous diffusion model prediction \hat{x}_0^t and use \hat{x}_0^t as the next step's prediction \hat{x}_0^{t-1} .

B Additional Details of Ablation Studies

Figure 1 showcases the baseline mentioned in Table 6 of the section "Effectiveness of Different Feature Processing" within the 5.4 Ablation Studies. In this baseline, we have substituted the feature processing operation from DSG+ (baseline) into MAT-A2G, providing a foundation for comparison with other methods.

Figure 2 presents the AdaLN-Zero, as mentioned in Table 6 from the "Effectiveness of Different Feature Processing" section within the 5.4 Ablation Studies.

C Compared to DDIM

As shown in Table 2, we observed that directly using DDIM [5] to accelerate the inference process of MDT-A2G does not yield satisfactory results. We still need to compute the reverse process for each step during the acceleration phase, e.g., with an acceleration ratio of 1:20. In the i -th step, we obtain \hat{x}_0^i through denoising network computation, and then use Eq. 8 to calculate the predicted value for the next 20 steps.

D More details about feature processing

Emotion: The BEAT dataset's eight emotions are encoded as one-hot vectors and transformed by a linear layer into the emotion feature. We transform these labels into a continuous feature space through a linear layer.

We don't need to ensure consistent lengths for audio and text, only the same number of frames. Assuming a batch size of 150 and 300 frames, we have $\hat{x}_a \in \mathbb{R}^{150 \times 300 \times 1131}$ and $\hat{x}_{txt} \in \mathbb{R}^{150 \times 300 \times 301}$. This allows us to concatenate \hat{x}_a and \hat{x}_{txt} along the last dimension.

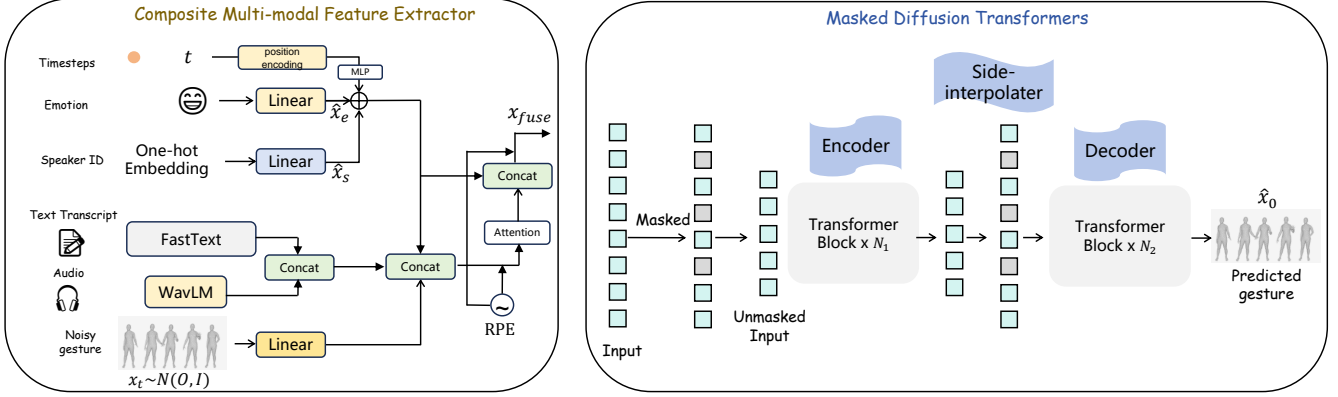


Figure 1: The baseline mentioned in Table 6 of the section "Effectiveness of Different Feature Processing" within the 5.4 Ablation Studies. RPE is Relative Position Encoding [6].

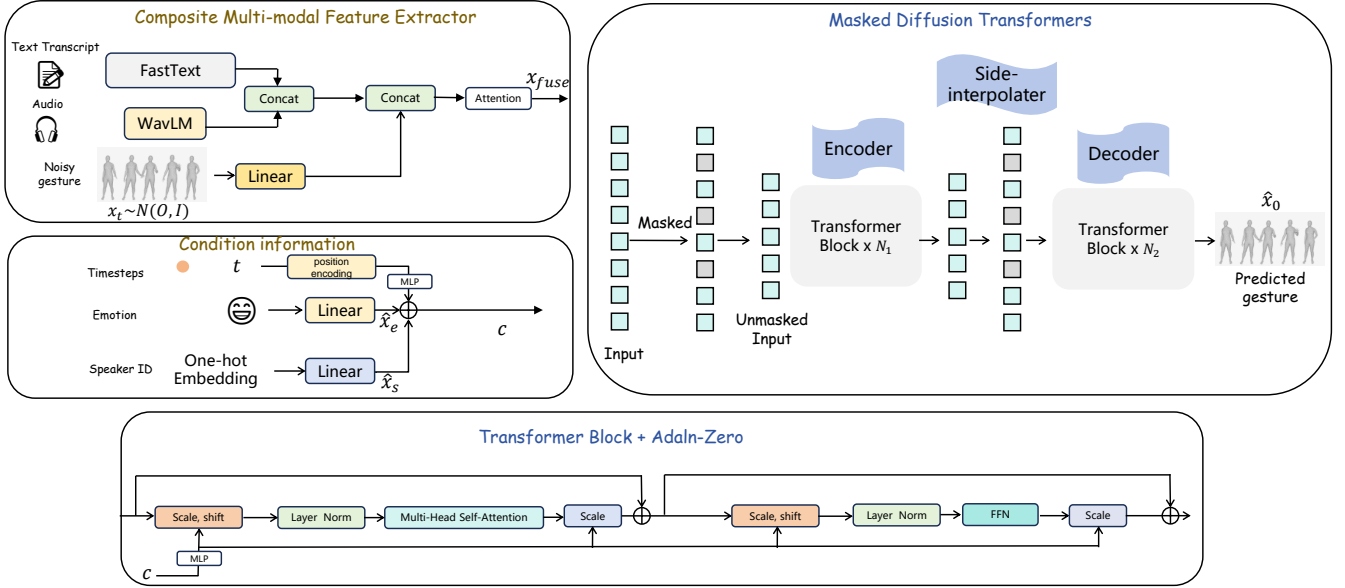


Figure 2: We have integrated the AdaLN-Zero [4] module into MDT-A2G, replacing the SEAFusion.

We generate multiple short sequences and then concatenate them, using interpolation and other post-processing techniques to ensure smooth transitions between segments.

E Model structure

Table 3: Network configurations of MDT-A2G models. N_2 is the number of decoders. The parameters and FLOs are measured during inference.

Size	Layers	N_2	Dim.	Head Num.	Param. (M)	FLOs (G)
Network configurations of MDT-A2G models.						
TS	3	1	384	6	3.43	0.9
S	5	2	384	6	5.01	1.4
B	8	2	384	6	8.17	2.3
L	12	4	512	6	15.4	4.5
Network configurations of DSG+ baselines.						
DSG+	8	-	384	6	7.68	2.2

References

- [1] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. 2022. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv*

- preprint arXiv:2201.06503* (2022).
- [2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
 - [3] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. 2023. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321* (2023).
 - [4] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.
 - [5] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
 - [6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).