SMooDi:Stylized Motion Diffusion Model

https://neuvi.github.io/SMooDi/

Lei Zhong[1], YiMing Xie[1], Varun Jampani[2], Deqing Sun[3], Huaizu Jiang[1]

[1]Northeastern University

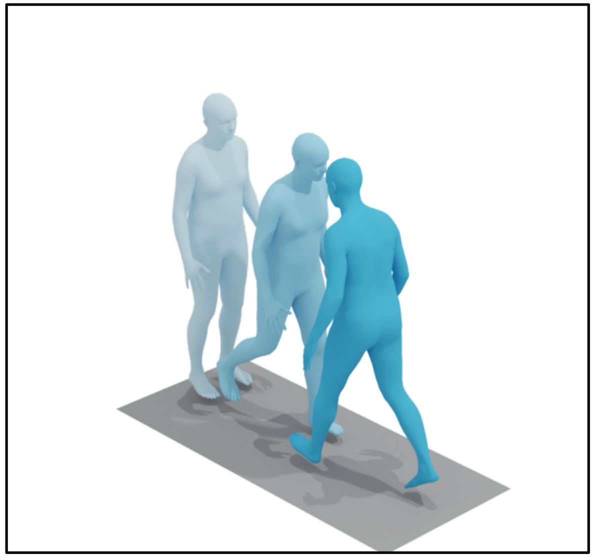
[2]Stability AI

[3]Google Research

Motivation

Text2Motion

Content Text:
*A person **walks** forward and then **sits** down.*

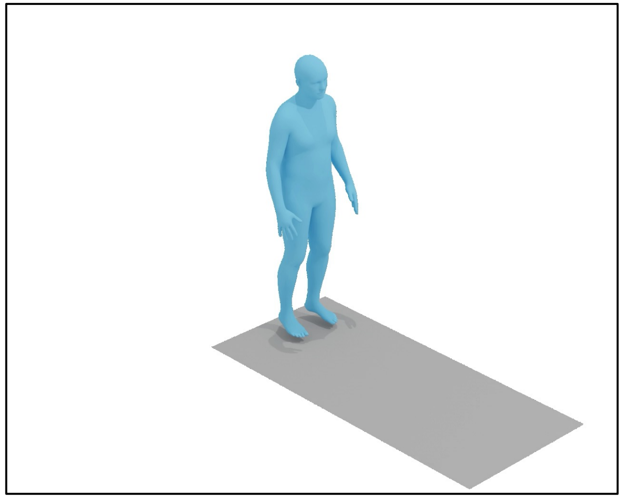


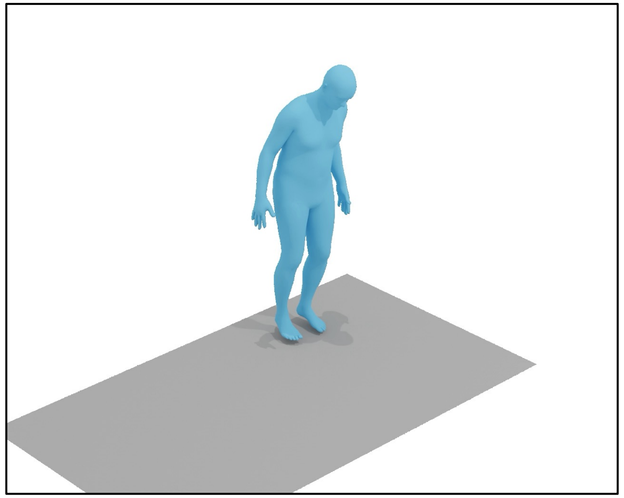
Text2Motion

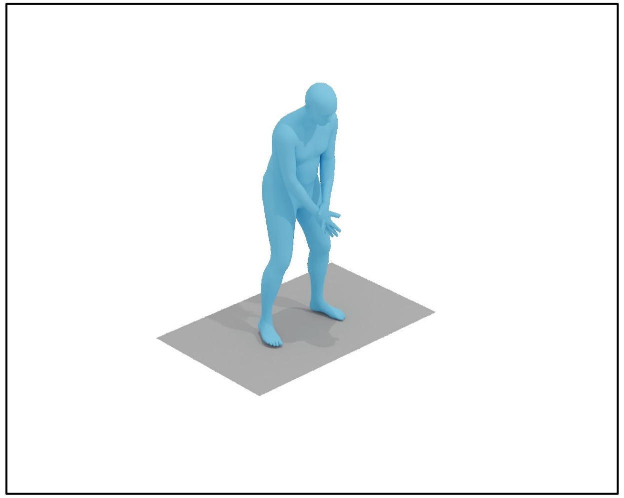
Text2Motion primarily focuses on translating content text into corresponding motions without considering the motion style.

HumanML3D dataset

The HumanML3D dataset contains ***diverse motion content*** but limited motion styles.

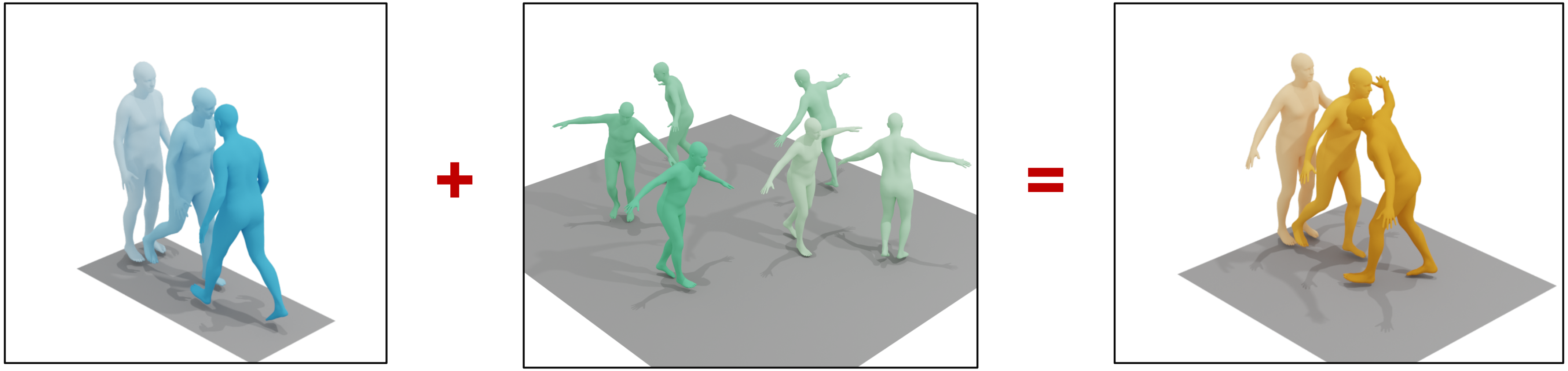






Motion Style Transfer

Giving a **content text** and a **style motion**, **MST** aim to generate a **stylized motion** that adheres to both content and style constraints

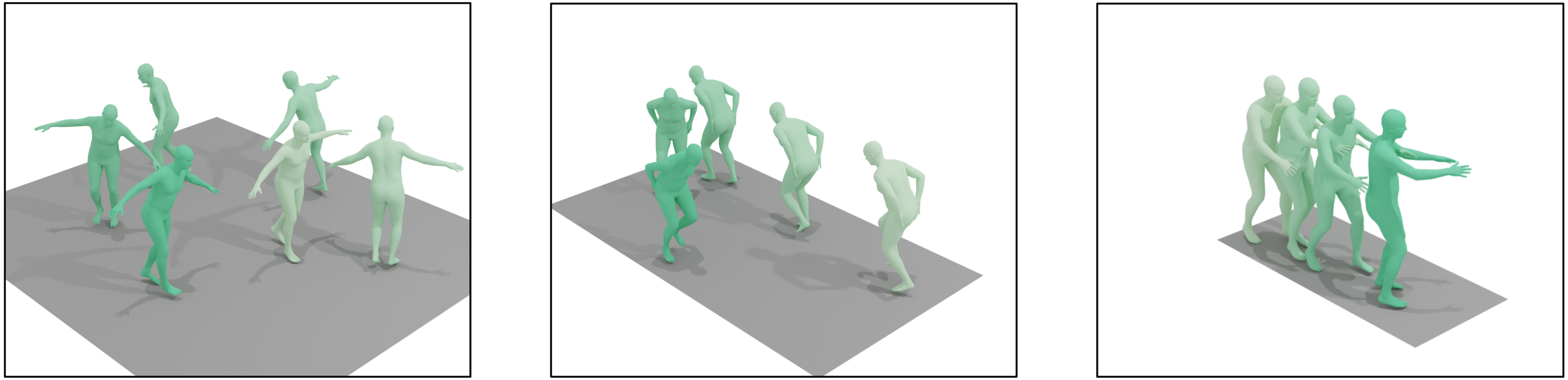


Content Motion　　Style Motion　　Stylized Motion

* MST means motion style transfer.

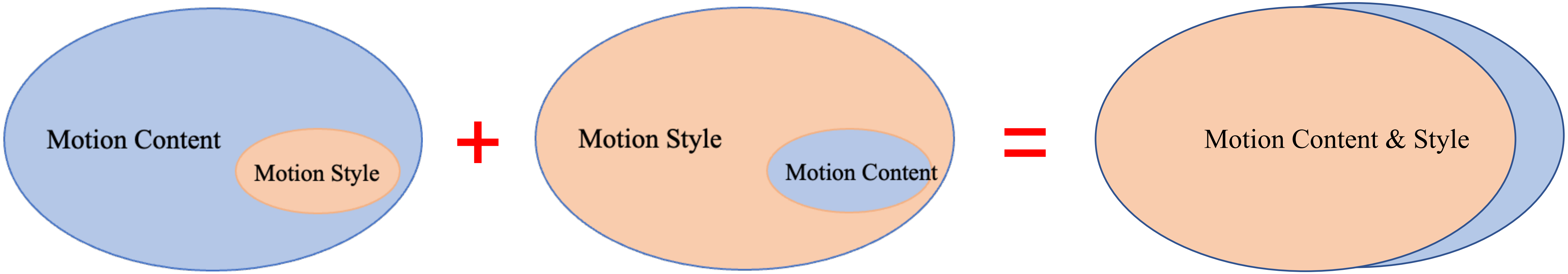
100STYLE dataset

The 100STYLE dataset contains up to 100 motion styles but only includes ***locomotion-related*** motion content.



Motivation

Could we apply **locomotion-style** to the existing **Text2Motion** model?

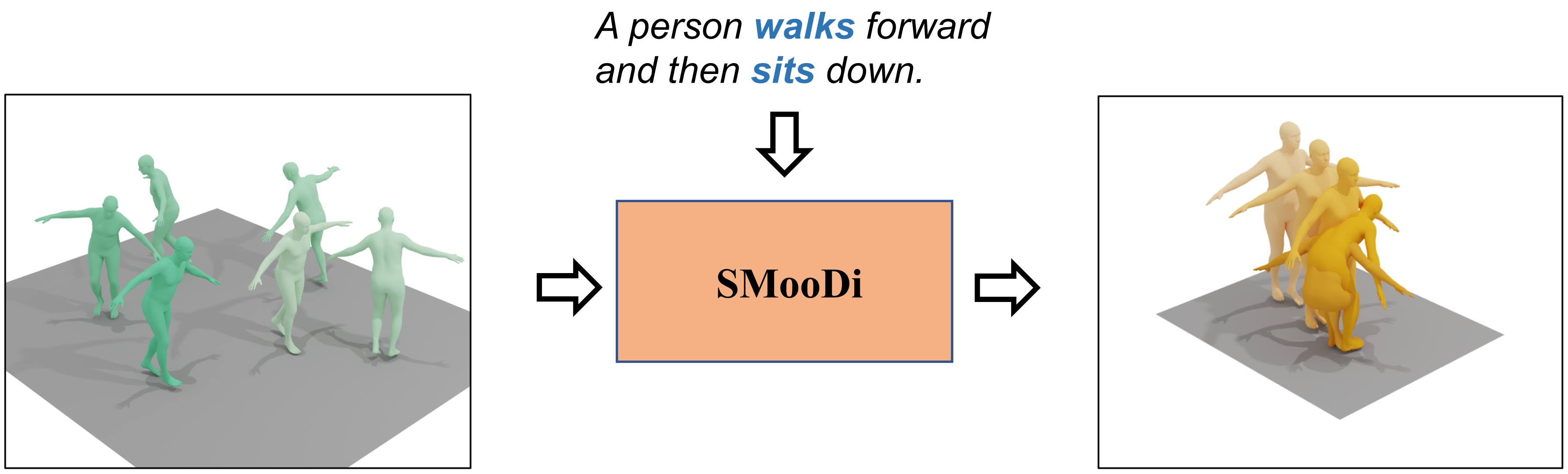


Text2Motion Model *100STYLE dataset* *Stylized Motion Generation*

Stylized Text2Motion

Stylized Text2Motion

Giving a content text and a style motion, **SMooDi** can generate a stylized motion that adheres to both content and style constraints

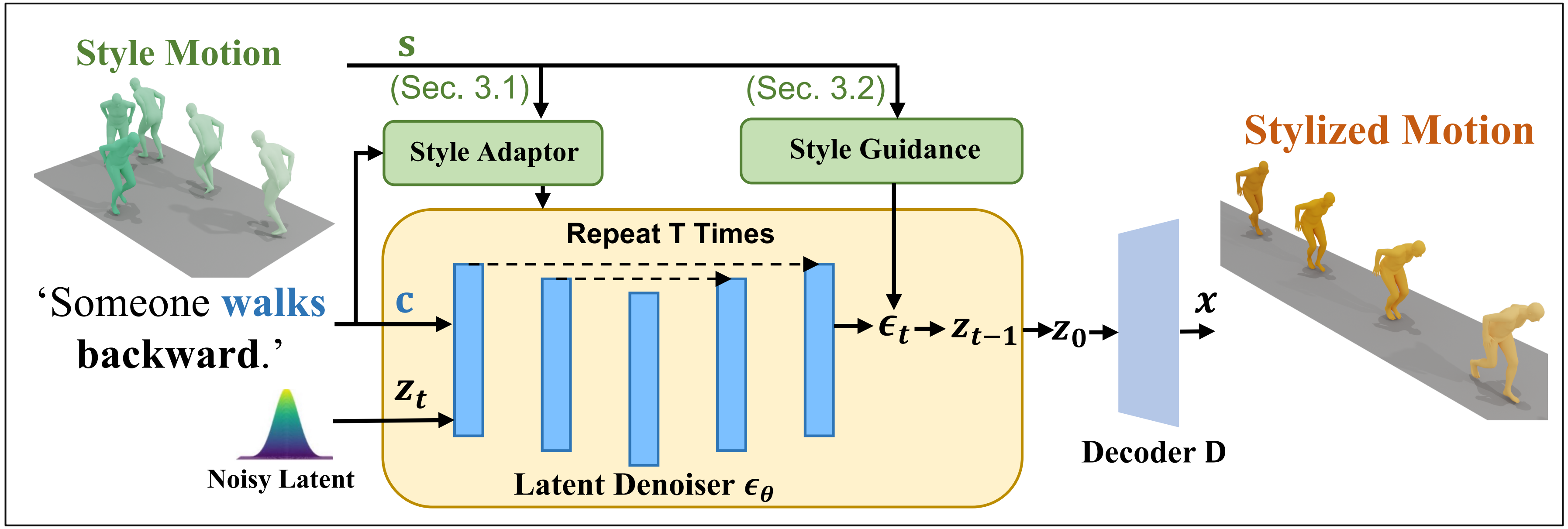


Style Motion

Generated Motion

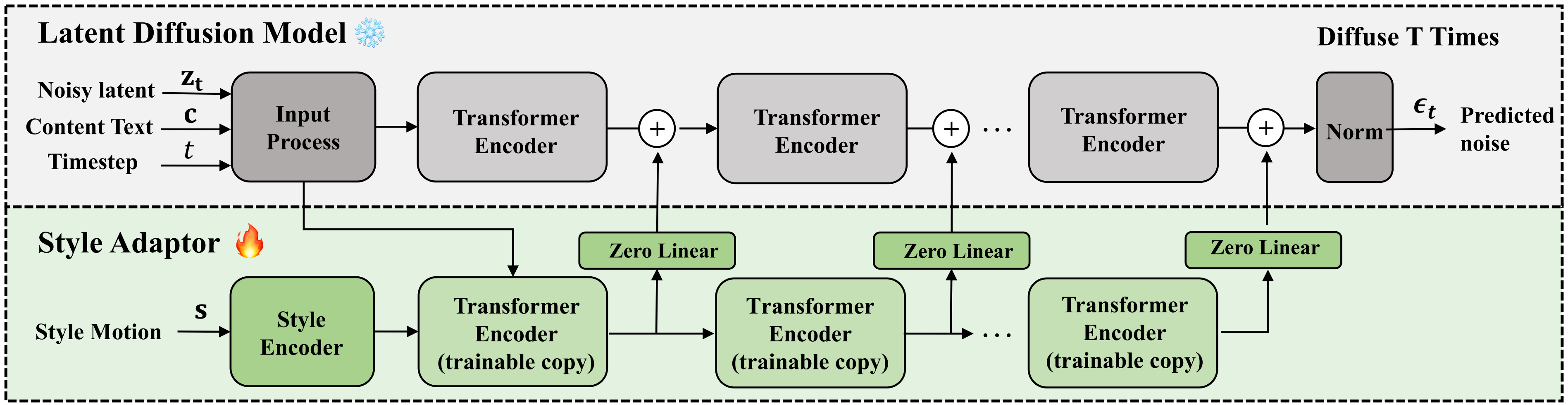
Stylized Text2Motion

Giving a content text and a style motion, **SMooDi** leverages **a style adaptor** and **style guidance** to enable stylized motion generation.



Stylized Text2Motion

Style Adaptor is a trainable copy of the Transformer encoder in the motion diffusion model to learn to enforce the style constraints.



Stylized Text2Motion

Pseudo Code

Algorithm 1 SMooDi's inference

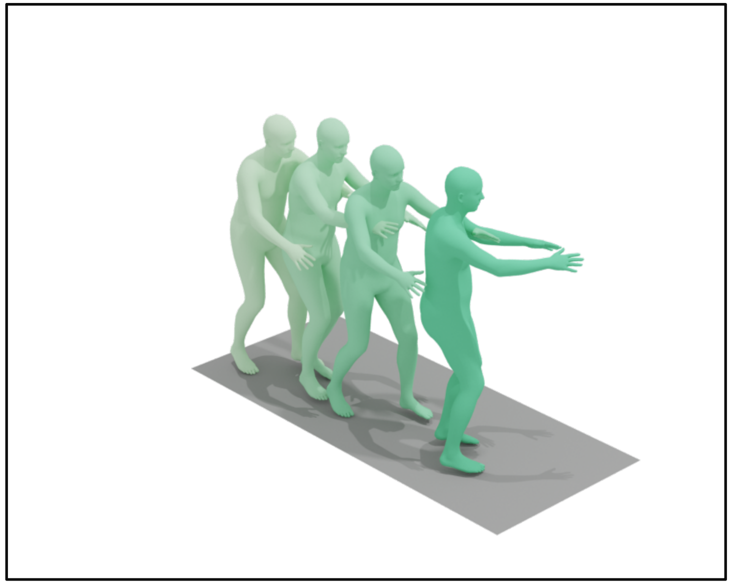
Require: A motion diffusion model M with parameters θ_M, a style adaptor model A with parameters θ_A, style motion sequence $\boldsymbol{s}$ (if any), content texts $\boldsymbol{c}$ (if any).

$\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ # Sample from pure Gaussian distribution
for all t from T to 1 **do**
  $\{\boldsymbol{r}\} \leftarrow A(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{s}; \theta_A)$ # **Style Adaptor model**
  $\epsilon_t \leftarrow M(\boldsymbol{x}_t, t, \boldsymbol{c}, \{\boldsymbol{r}\}; \theta_M)$ # Model diffusion model
  for all k from 1 to K **do** # **Classifier-based style guidance**
    $\epsilon_t = \epsilon_t + \tau \nabla_{\boldsymbol{z}_t} G(\boldsymbol{z}_t, t, \mathbf{s})$
  end for
  $\boldsymbol{z}_{t-1} \sim \mathcal{S}(\boldsymbol{z}_t, \epsilon_t, t)$ # $S(\cdot, \cdot, \cdot)$ represents the DDIM sampling method [10].
end for
$\boldsymbol{x}_0 = \mathbf{D}(\boldsymbol{z}_0)$
return $\boldsymbol{x}_0$

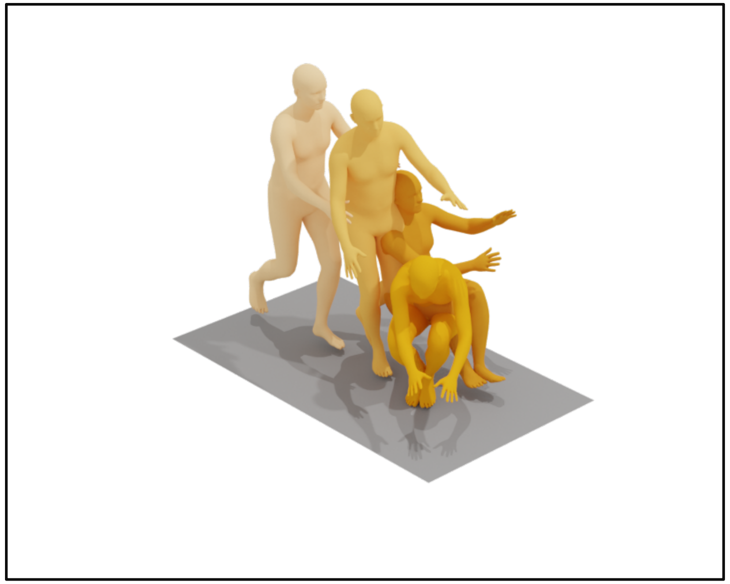
Stylized Text2Motion

Content Text:

*A person **walks** forward and then **sits** down.*



Style Motion

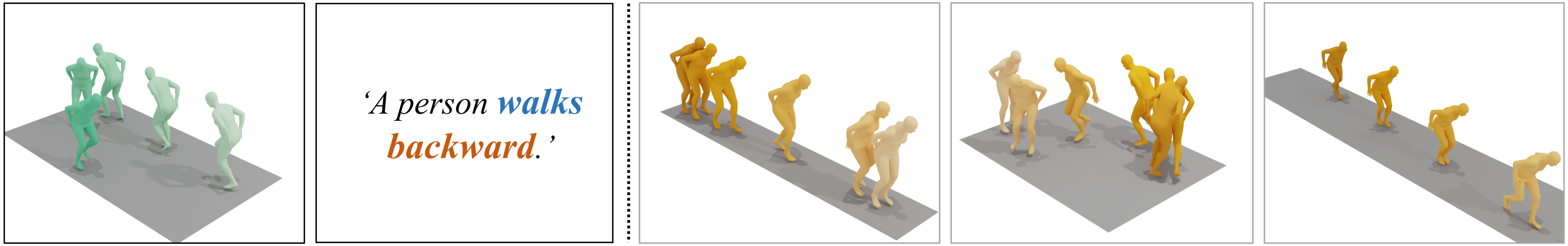


SMooDi

Stylized Text2Motion

- The straightforward baselines involve applying motion style transfer methods to the motion sequences generated by the text2motion model.

- **SMooDi** achieves **better** performance both quantity and quality.

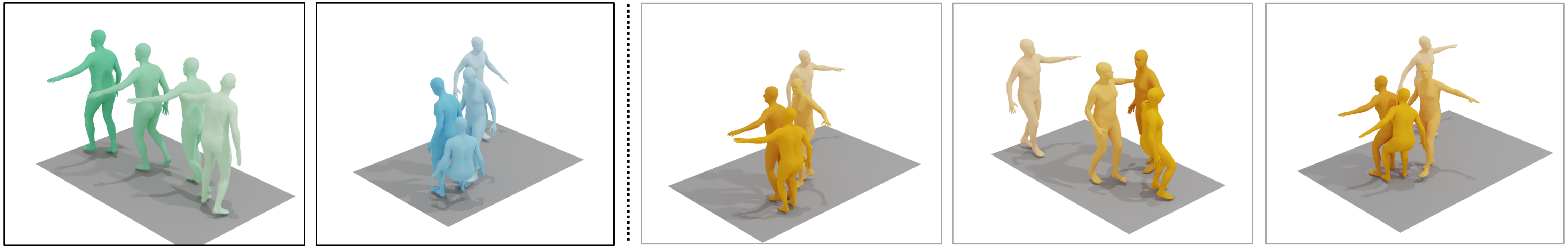


Style Motion | Content Text | (a) MLD+Motion Puzzle | (b) MLD+Aberman et al. | (c) Ours

Motion Style Transfer

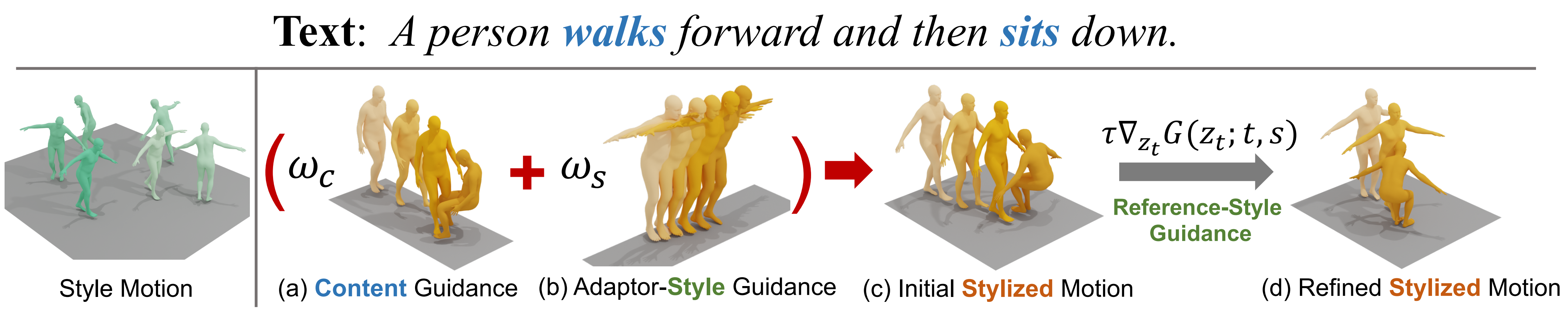
Stylized Text2Motion

Through DDIM inversion, **SMooDi** enables motion style transfer and achieves performance comparable to existing methods.



Style Motion | Content Motion | (d) Motion Puzzle | (e) Aberman et al. | (f) Ours

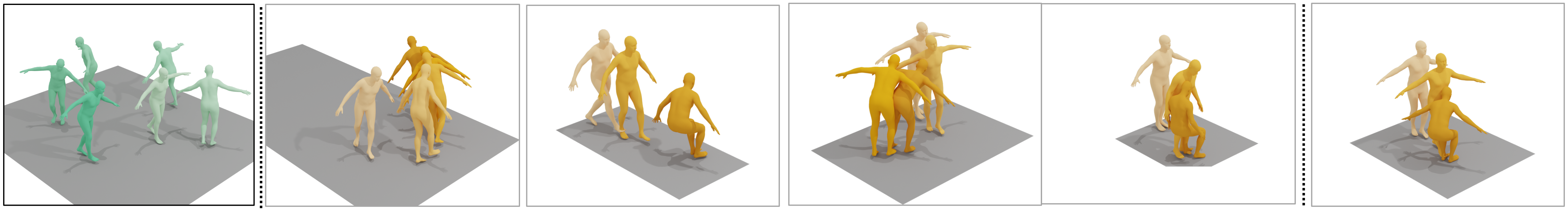
Visualize Style Guidance



Ablation Studies

Ablation Studies

Text: A person ***walks forward*** *and then* ***sits down****.*

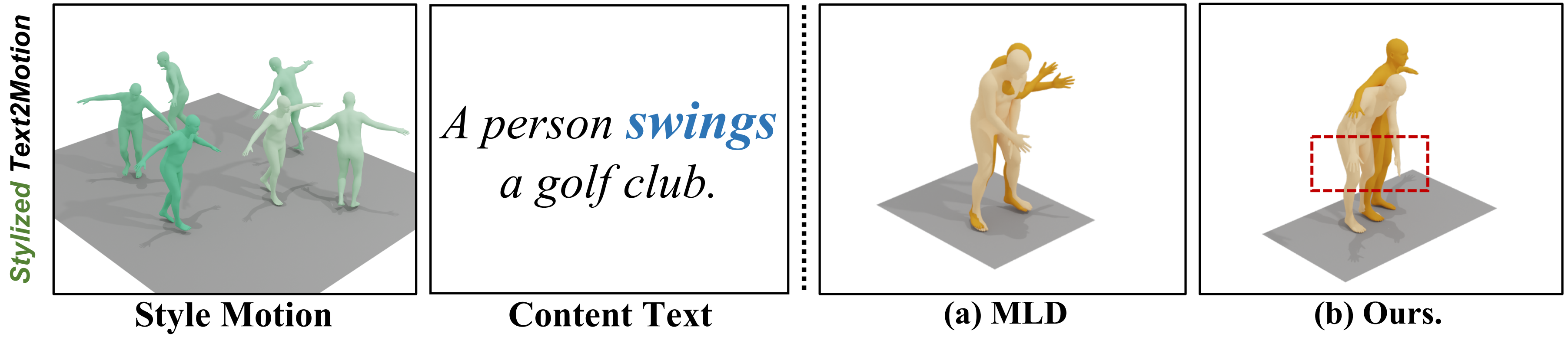


Style Motion | **(a) W/o L_{prior}** | **(b) W/o L_{cycle}** | **(c) W/o Adaptor** | **(d) W/o Ref-Guidance** | **(e) Full model**

Failure Cases

Failure Cases

When there are conflicts between the **content text** and the **style motion** in a specific body part, **SMooDi** may generate unrealistic motions.



Thanks