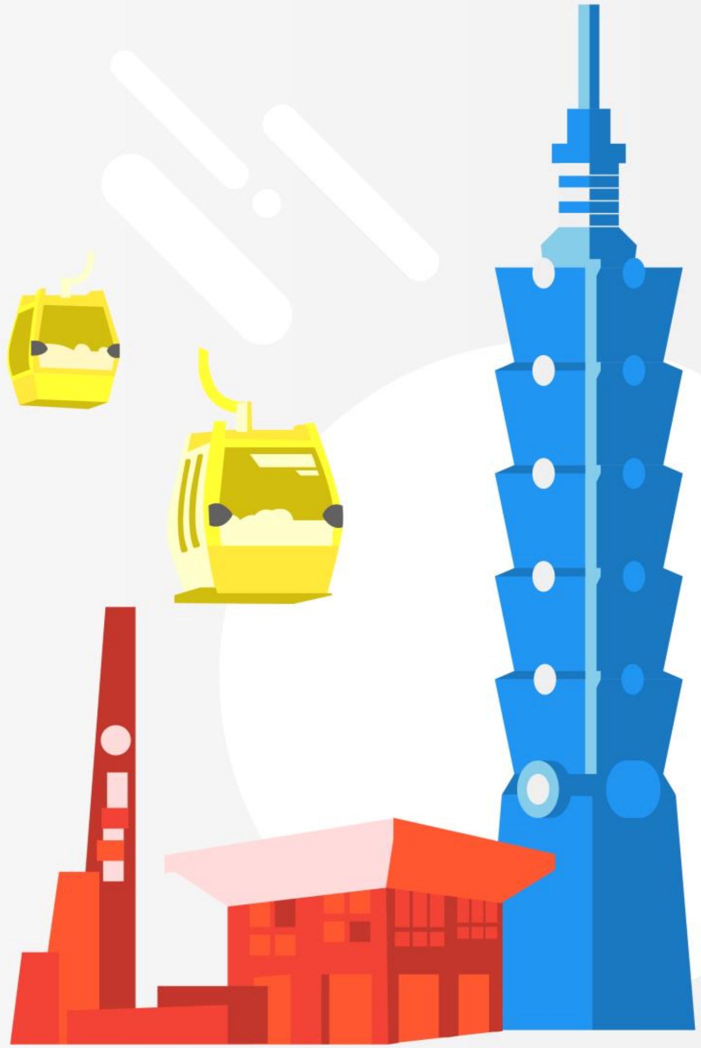
2025 TAIPEI

B2F:
End-to-End Body-to-Face Motion Generation with Style Reference

Bokyung Jang, Eunho Jung, Yoonsang Lee*

Hanyang University



Motivation

Humans perceive **face and body** as an integrated whole.
If a character's facial expressions do not align with its body motions,

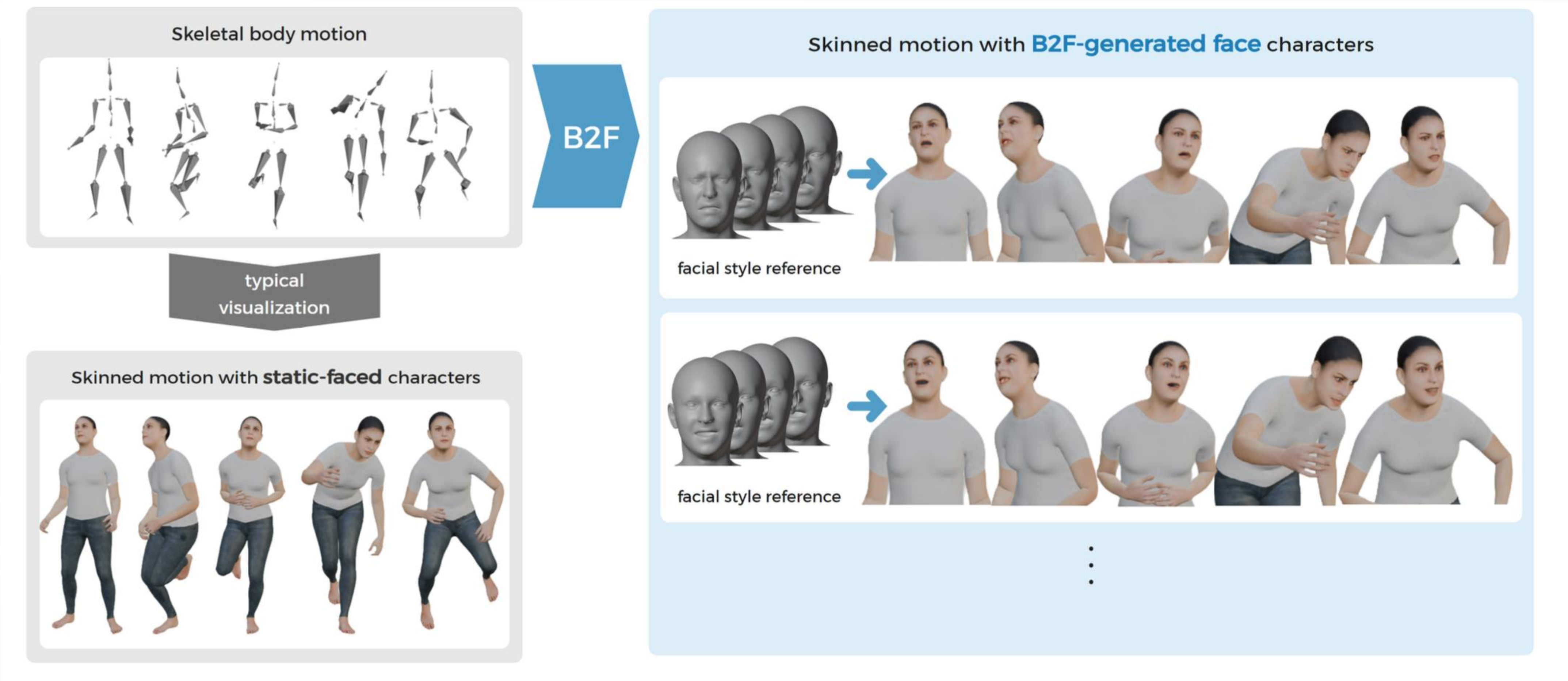
Xie et al. 2022

Ling et al. 2020

Could it **hinder the observer's perception and understanding?**

B2F

- We propose **B2F**, a method that **generates facial motions aligned with body motions**, while allowing style control.
- This improves perception by reducing face–body mismatches and leads to coherent, expressive animation.

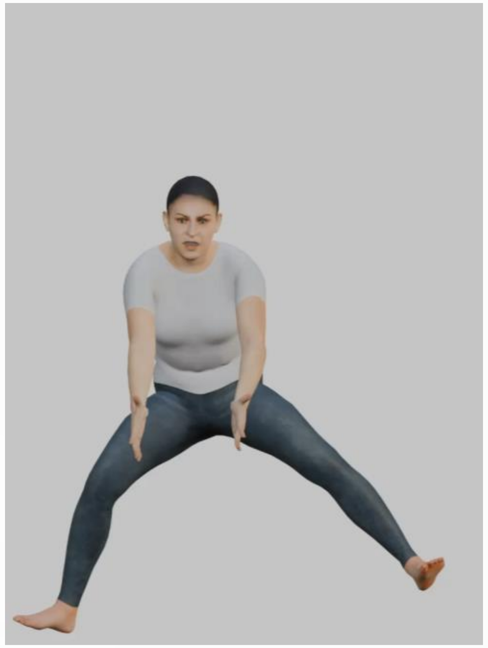



Static face

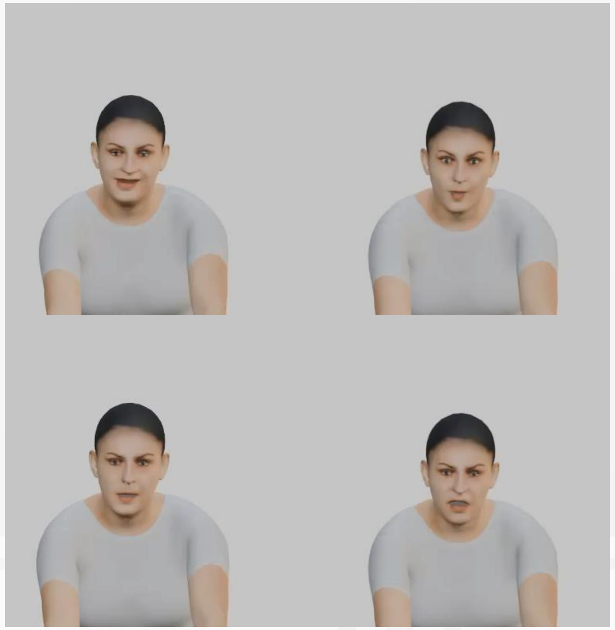
Our Approach

We assume that facial motion can be decomposed into **two factors**:

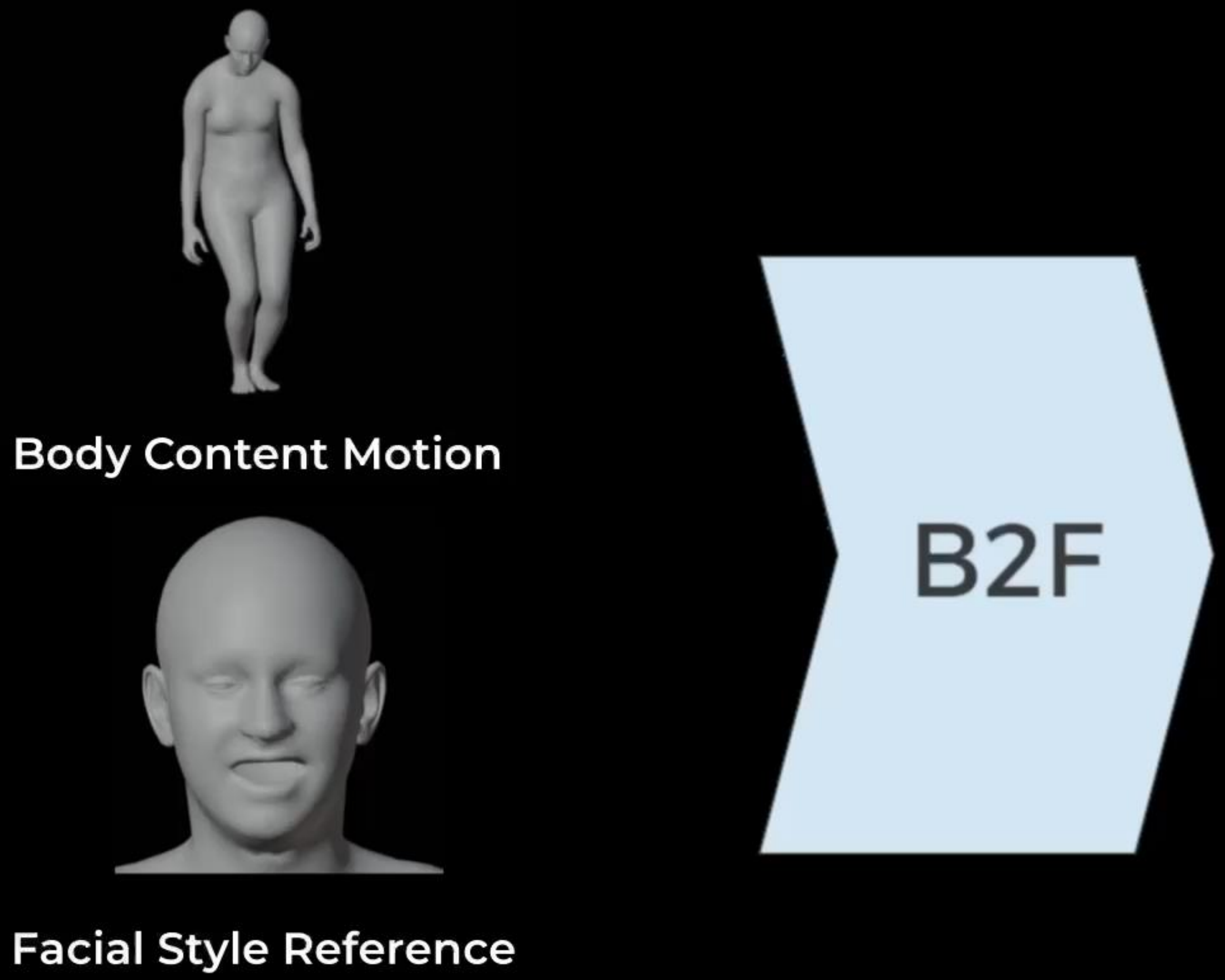
Content + **Style**



Reflects the context-dependent expressions that naturally accompany body movements

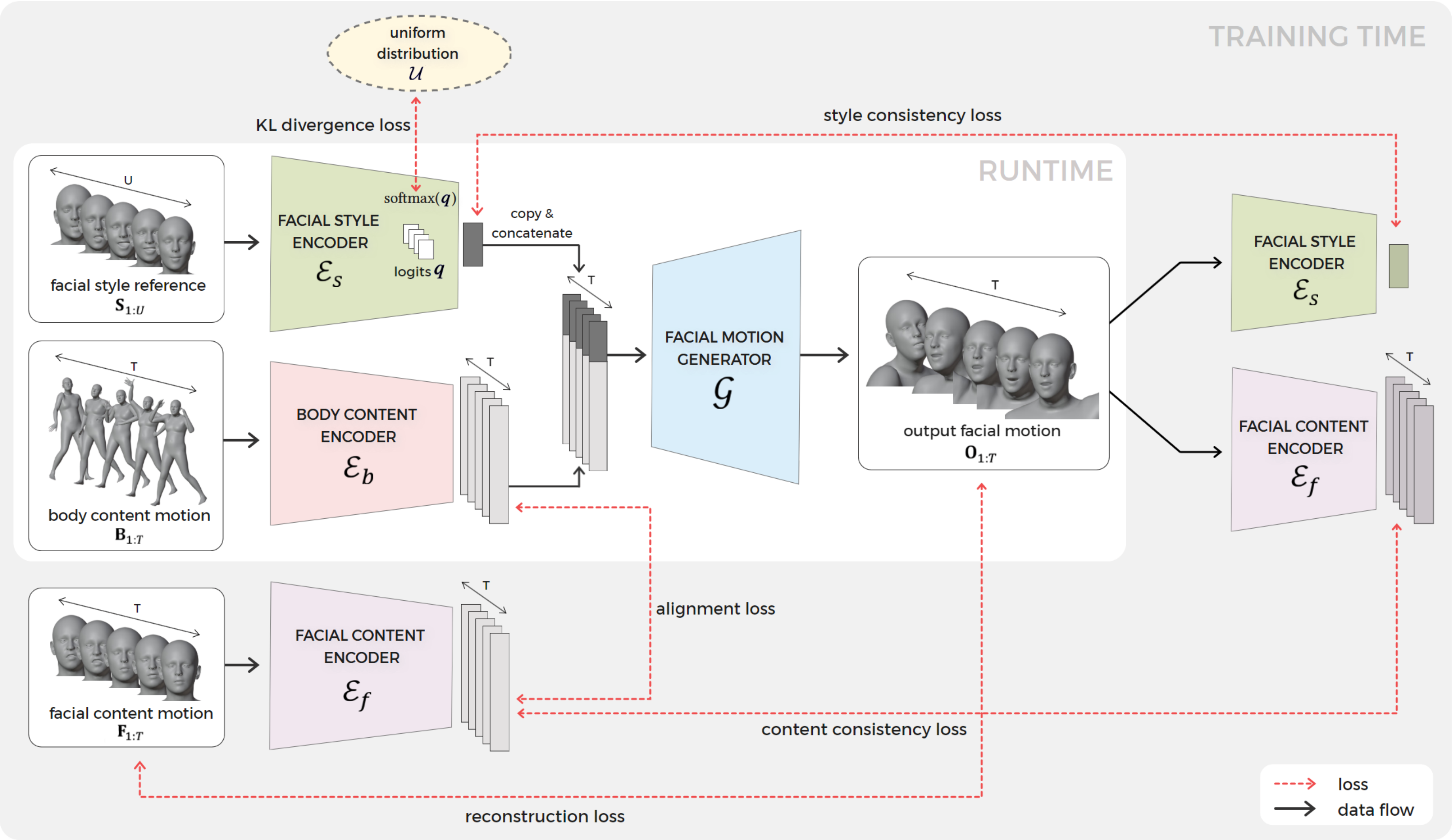


Captures the character's expressive tendency or emotional tone.


Body Content Motion
Facial Style Reference
B2F

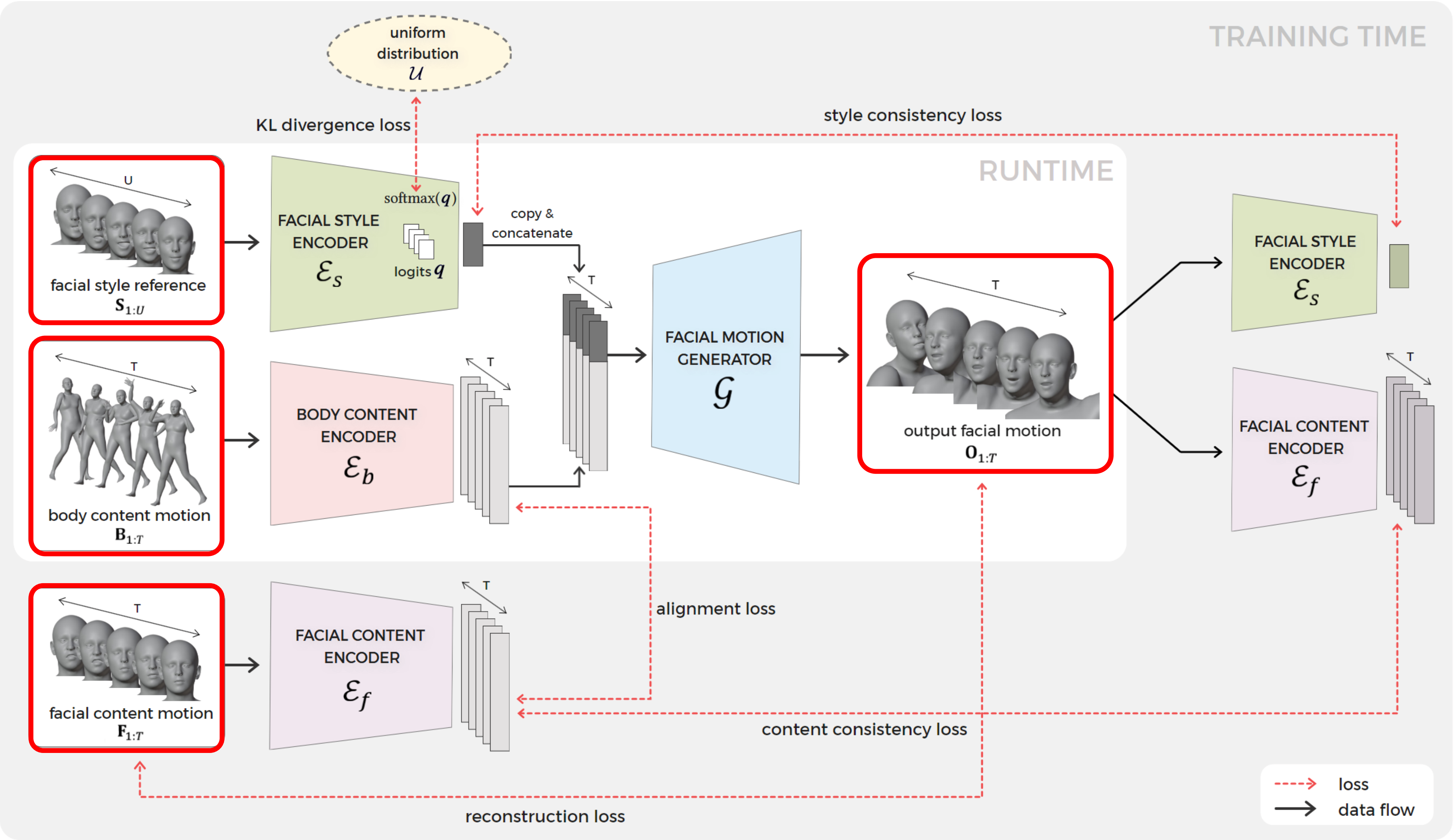
B2F Architecture

2025 TAIPEI

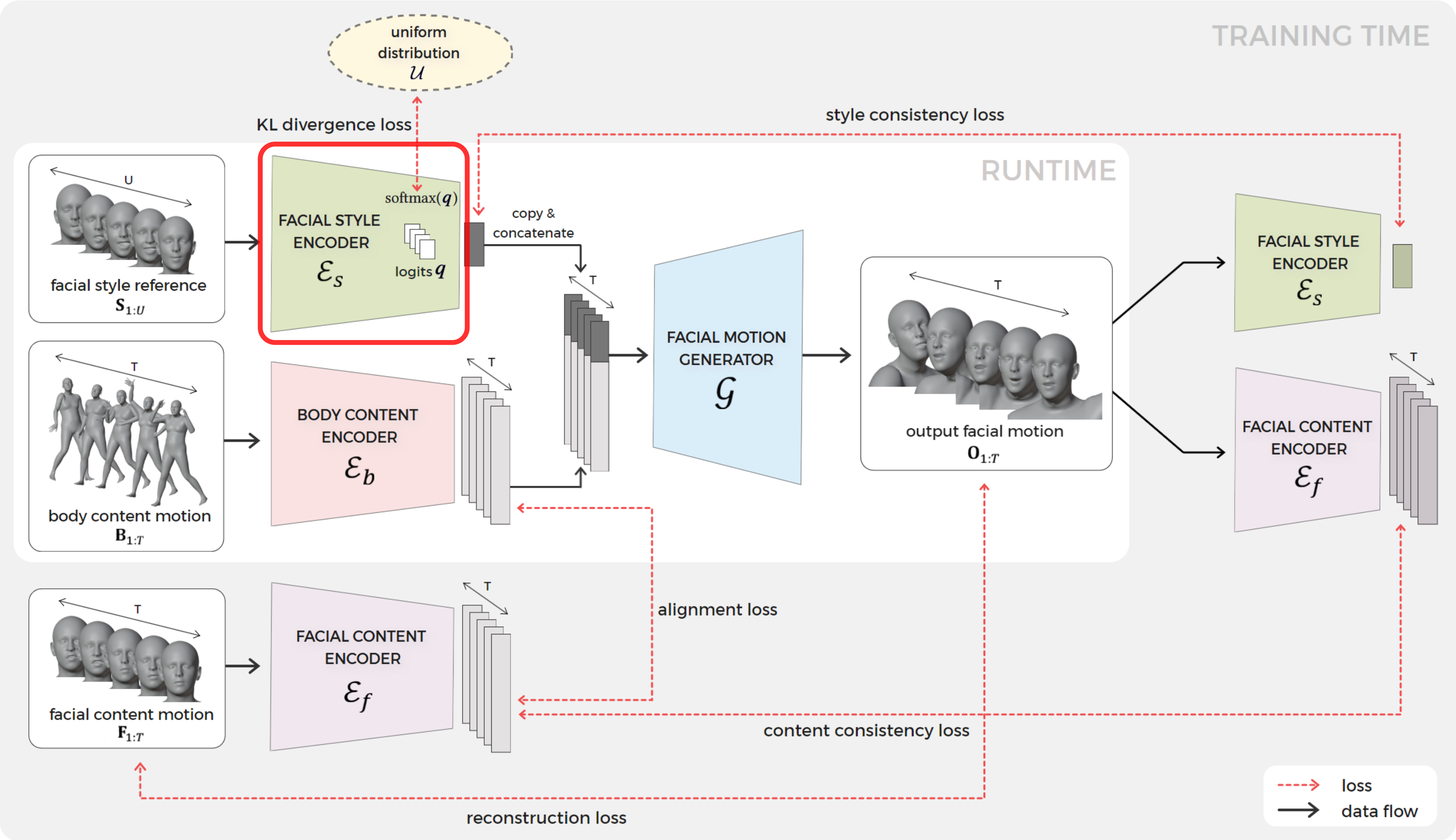


B2F Architecture

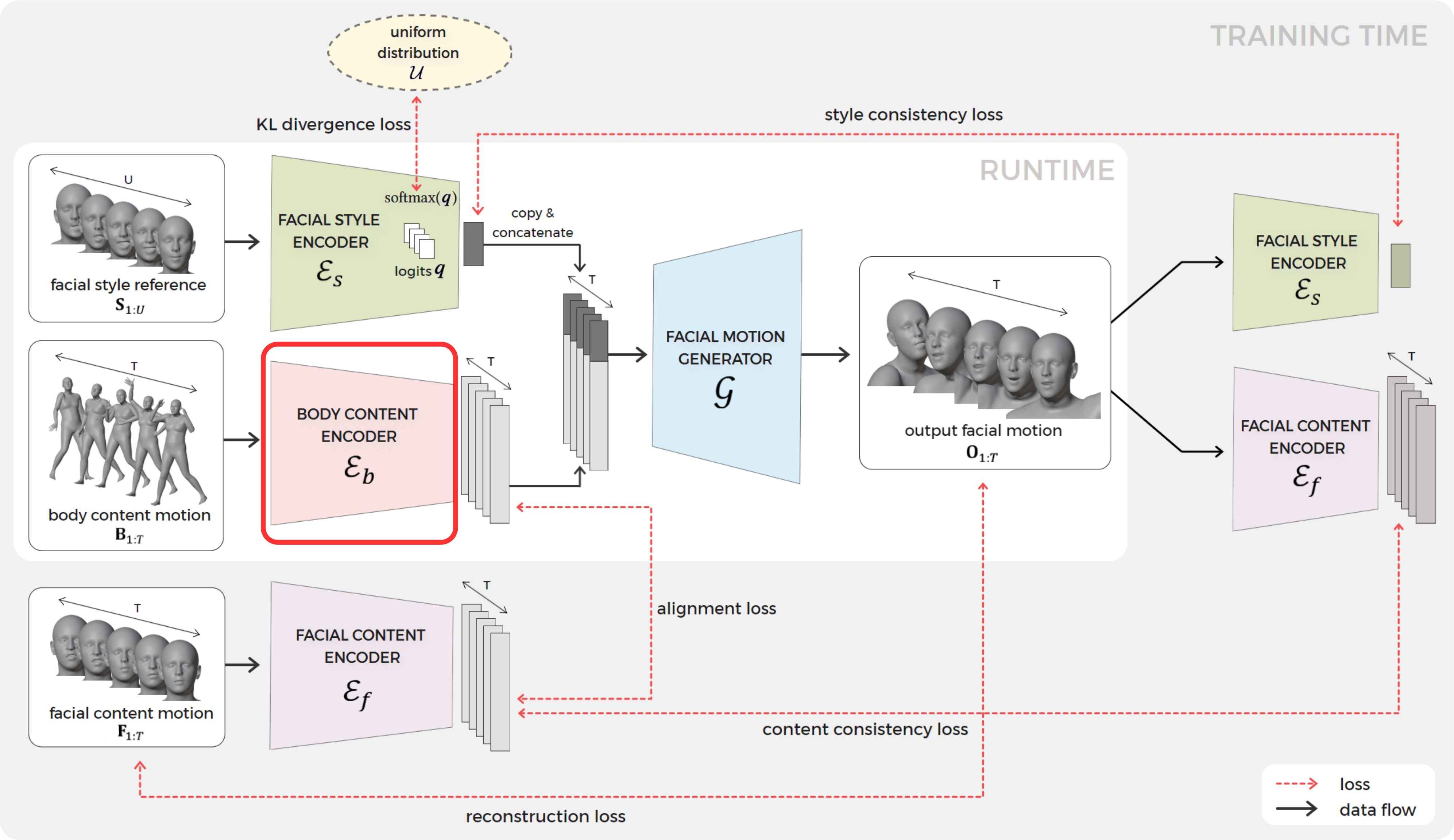
2025 TAIPEI



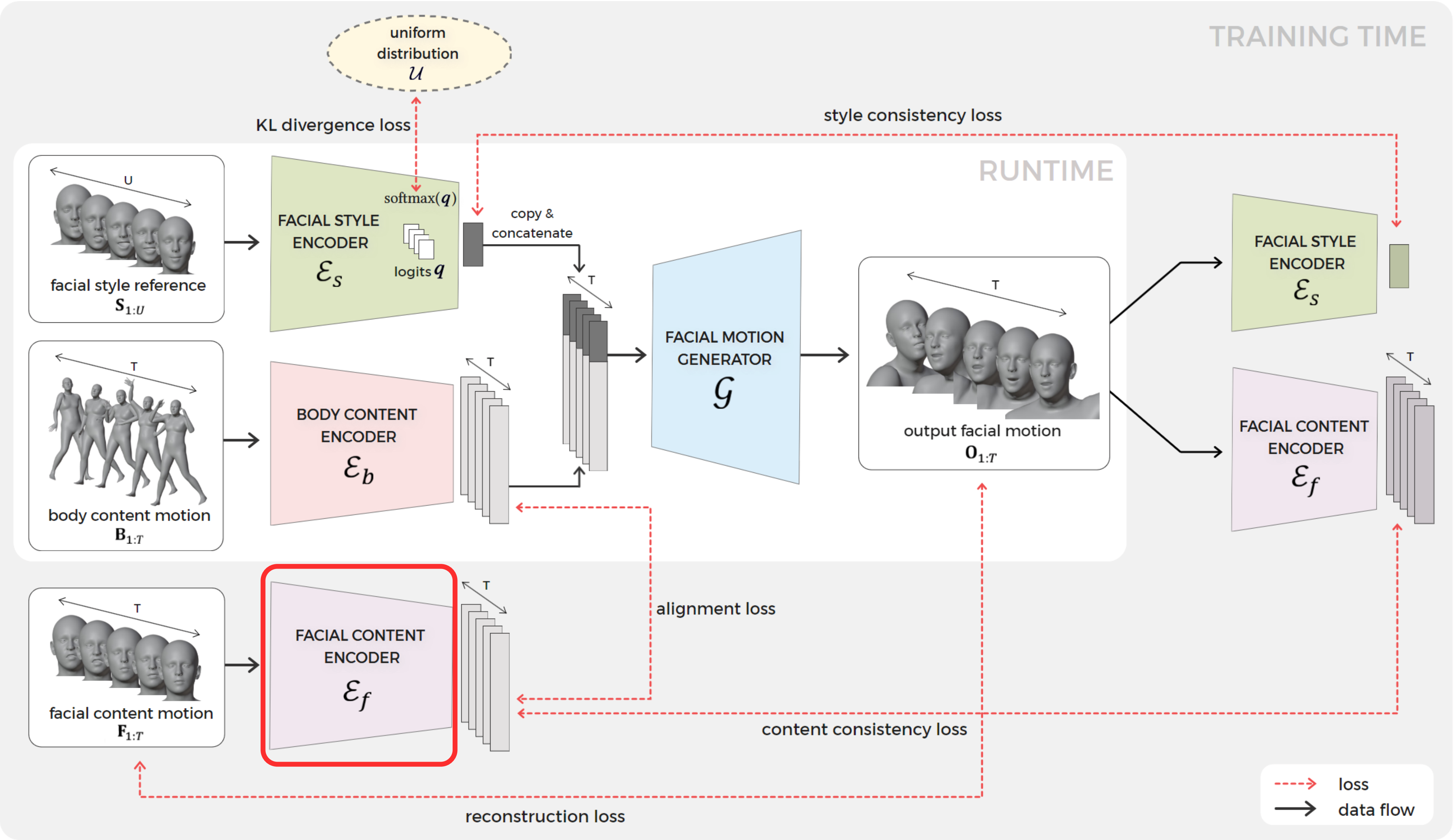
B2F Architecture



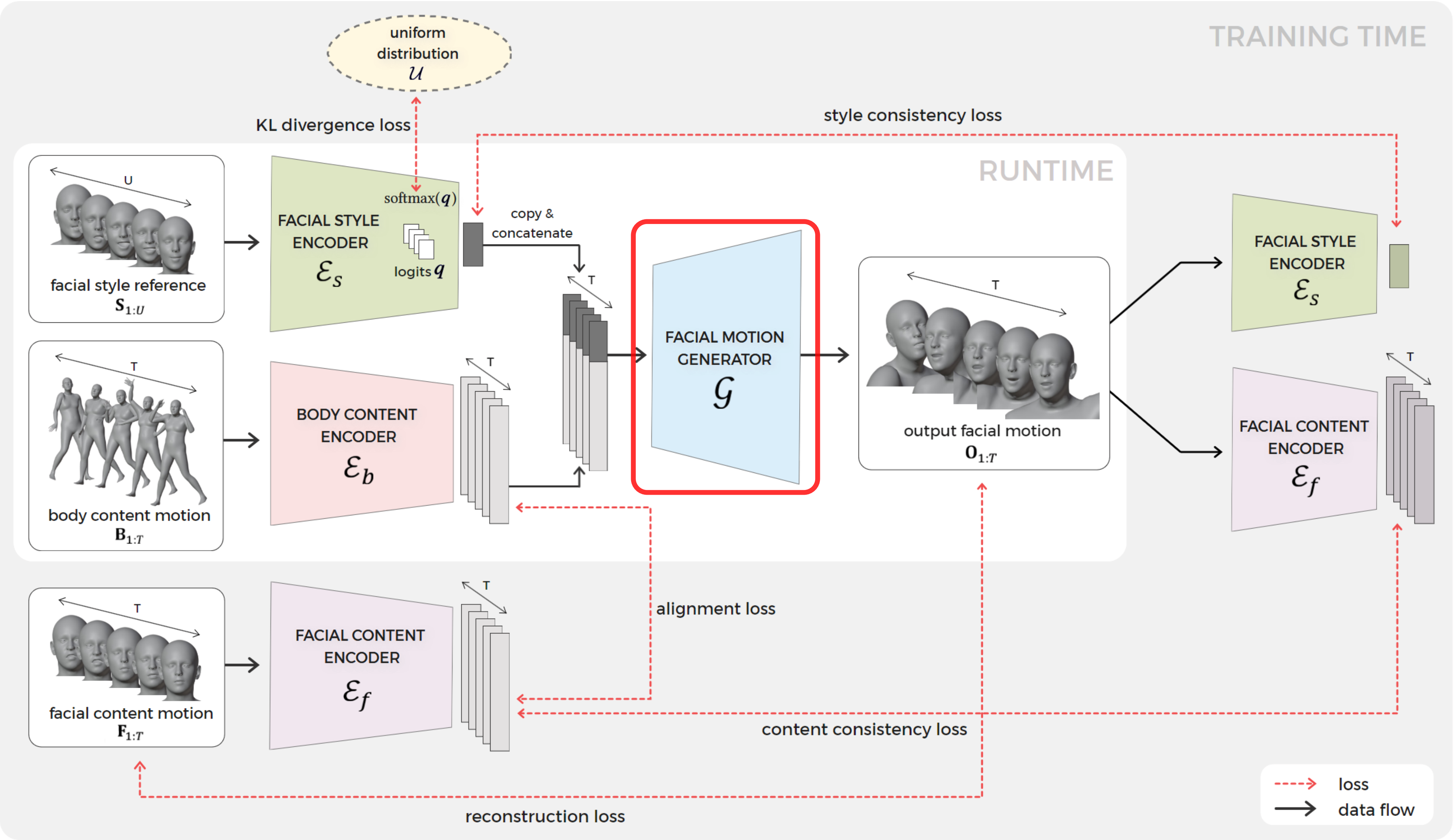
B2F Architecture



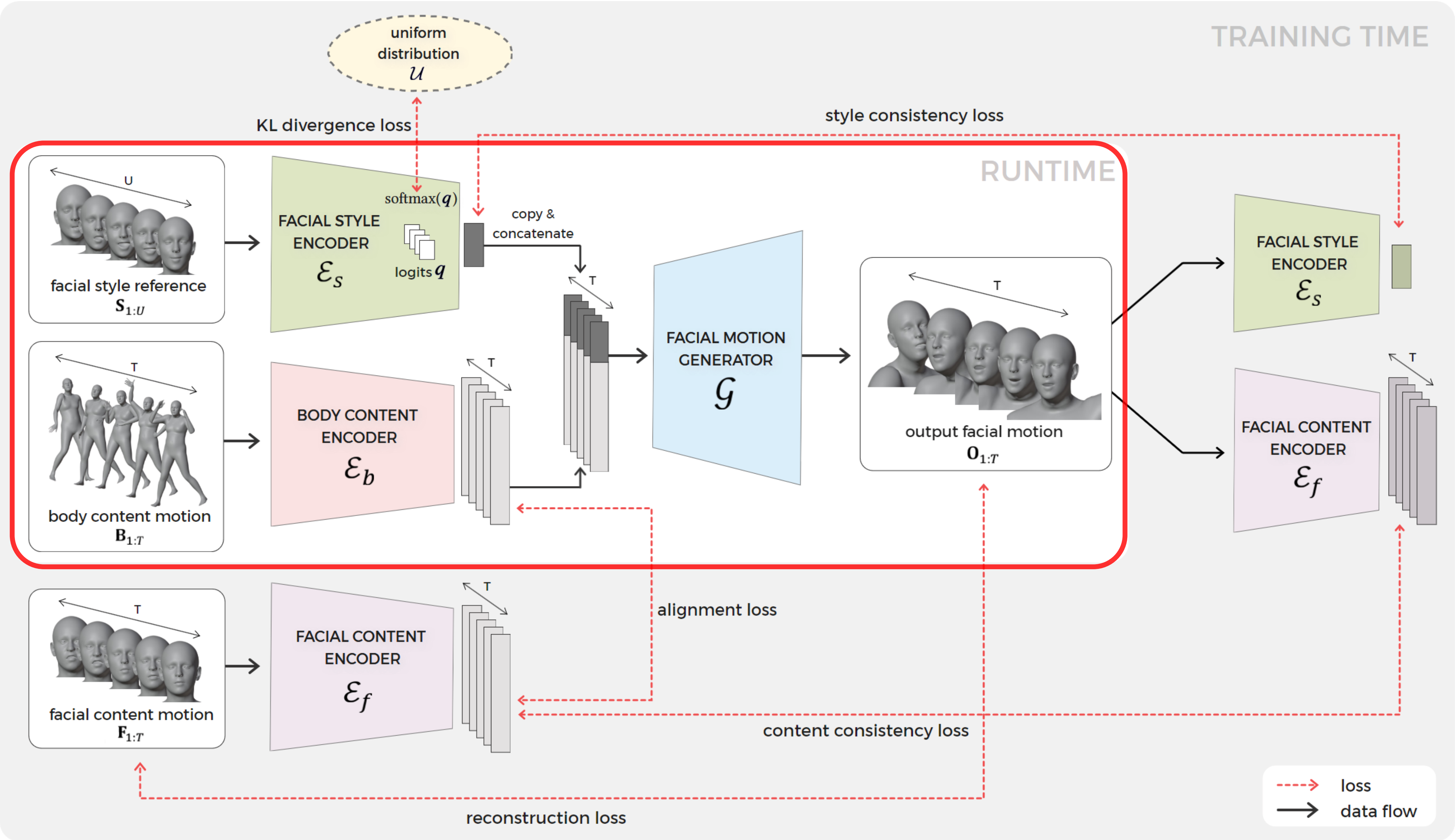
B2F Architecture



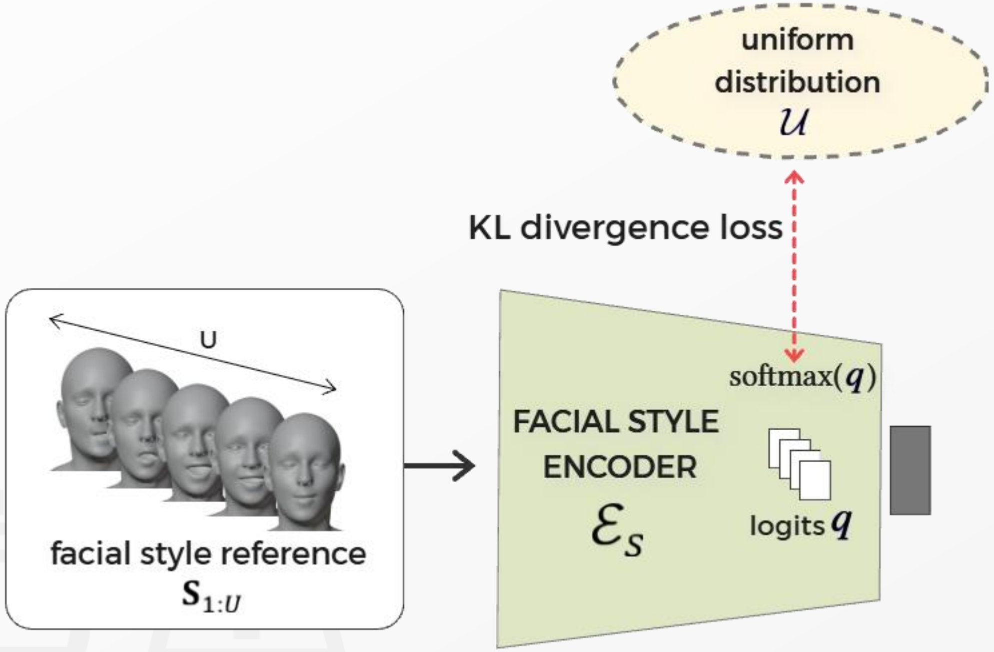
B2F Architecture



B2F Architecture

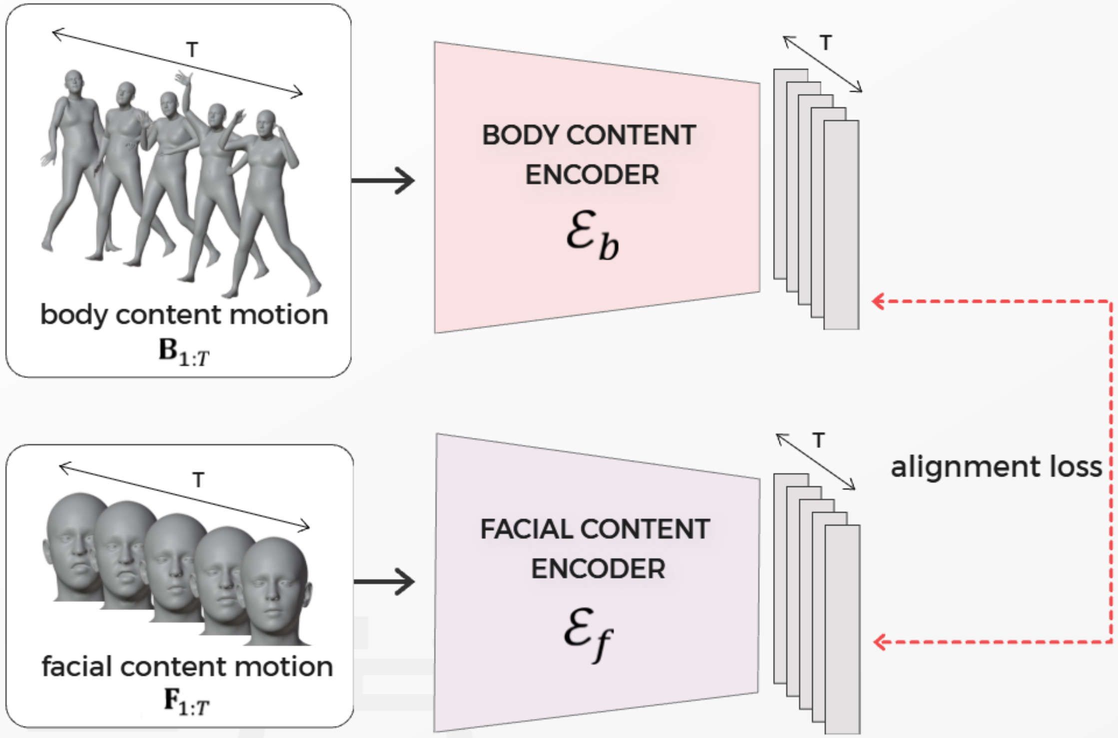


Facial Style Encoder



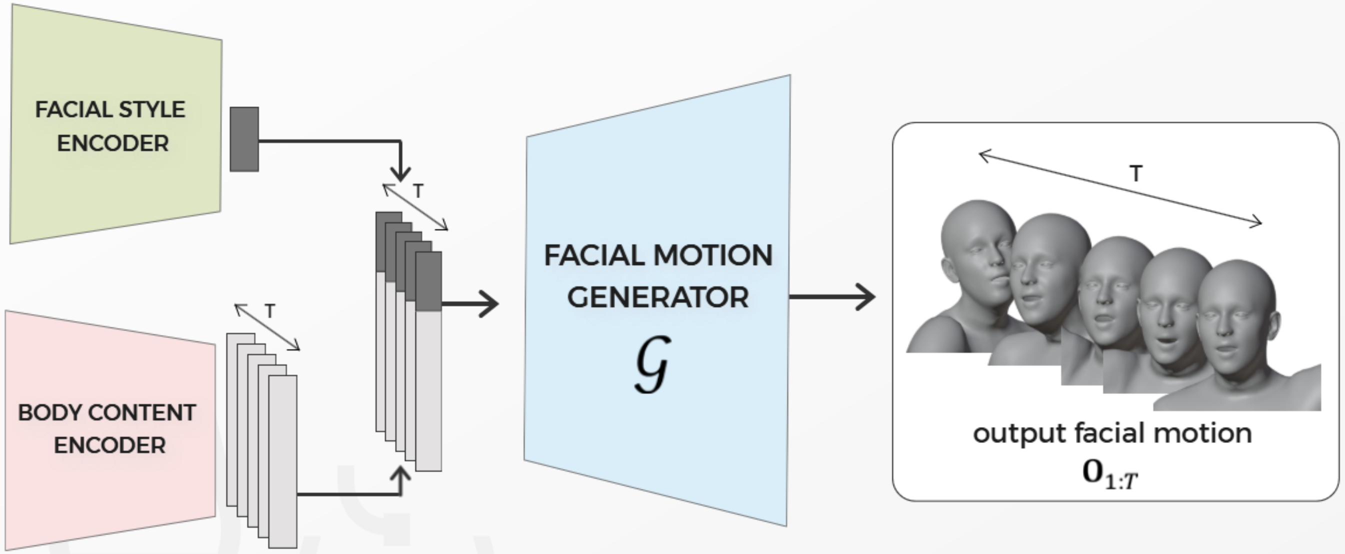
- Extracts a **style embedding** from a FLAME-based facial motion
- **Transformer encoder with Gumbel-Softmax** enables continuous and diverse style representation
- Achieves more **consistent and expressive results** than VAE/VQ-VAE

Body & Facial Content Encoder



- **Transformer encoder architecture** to **extract shared content** from body and face sequences.
- Trained with **co-temporal inputs** from the same clip to capture **common motion features.**
- **Facial Content Encoder** → used only during **training.**

Facial Motion Generator



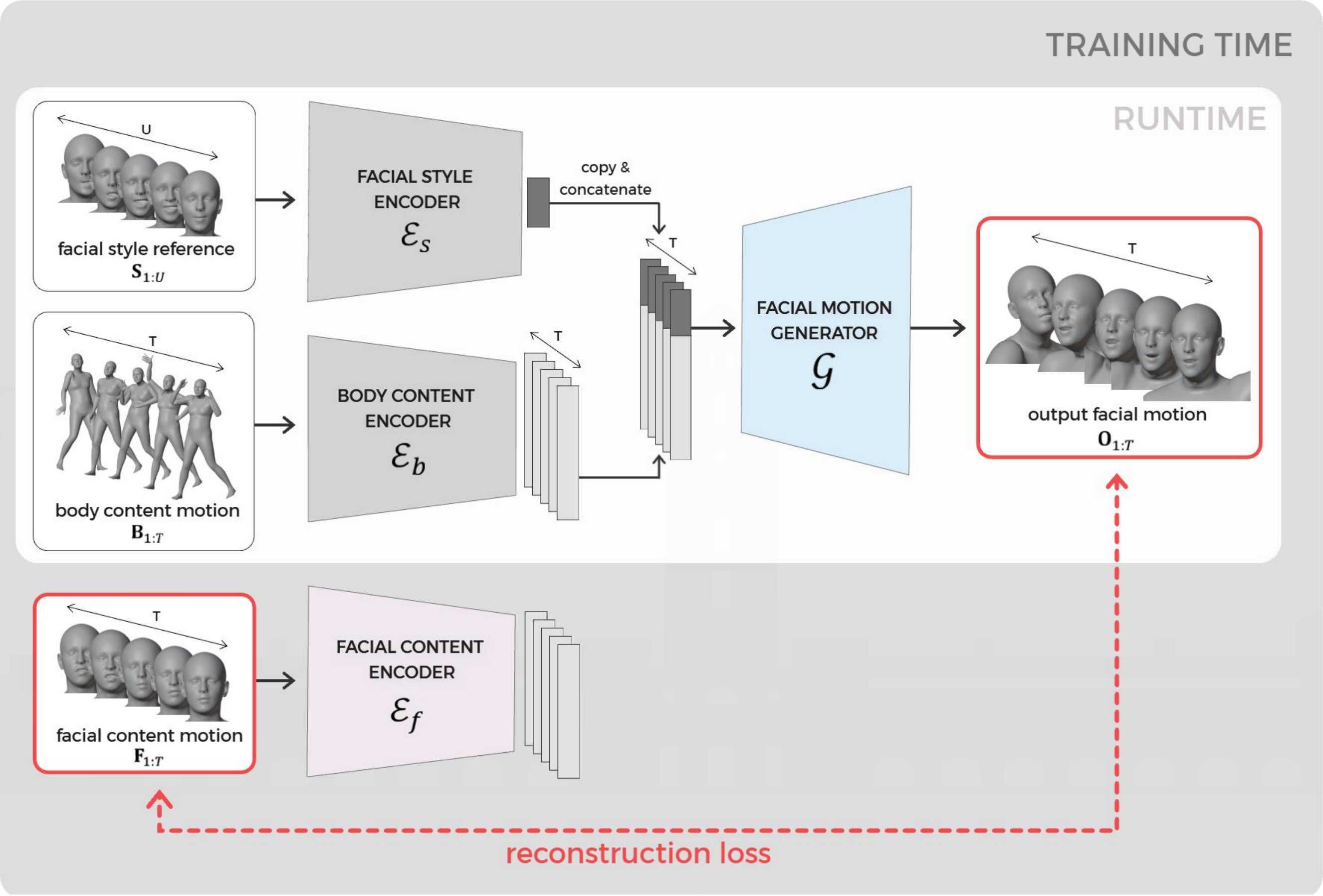
- **Inputs**: style + content embeddings
- **Repeated style** embedding **concatenated** with per-frame **content embeddings**.
- **Transformer decoder architecture** generates content- and style-aware facial motions

Training

$$\mathcal{L} = \lambda_1 \underline{\mathcal{L}_{\text{recon}}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{KL}} + \lambda_4 \mathcal{L}_{\text{consi}} + \lambda_5 \mathcal{L}_{\text{cross}}$$

- Reconstruction Loss

  MSE loss reconstructs original facial motion from **same-clip** content and style.



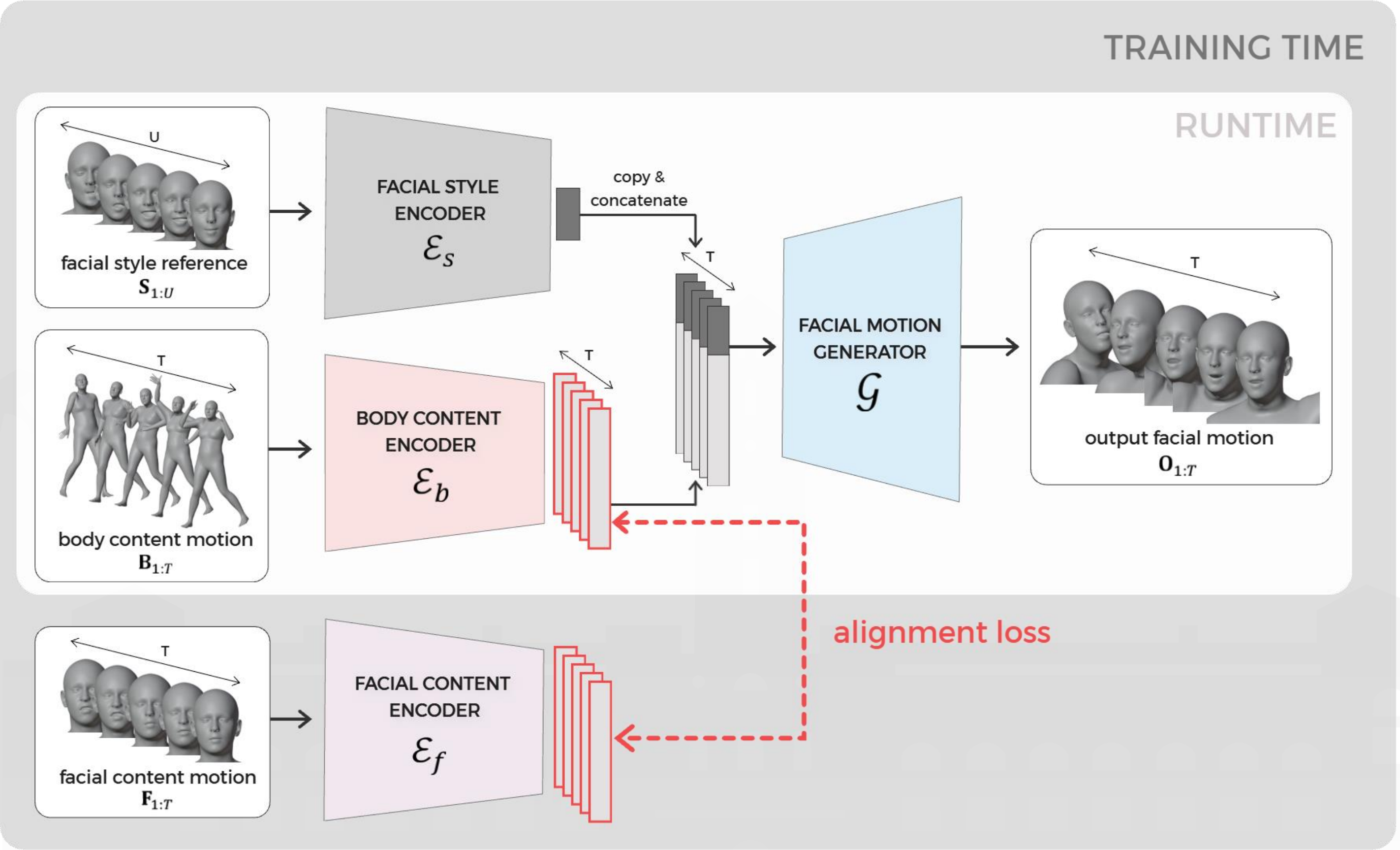
Training

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{KL}} + \lambda_4 \mathcal{L}_{\text{consi}} + \lambda_5 \mathcal{L}_{\text{cross}}$$

- Alignment Loss

  Cosine similarity aligns body and facial content embeddings for consistent content.

  Encourages **consistent motion representation** between the two encoders.



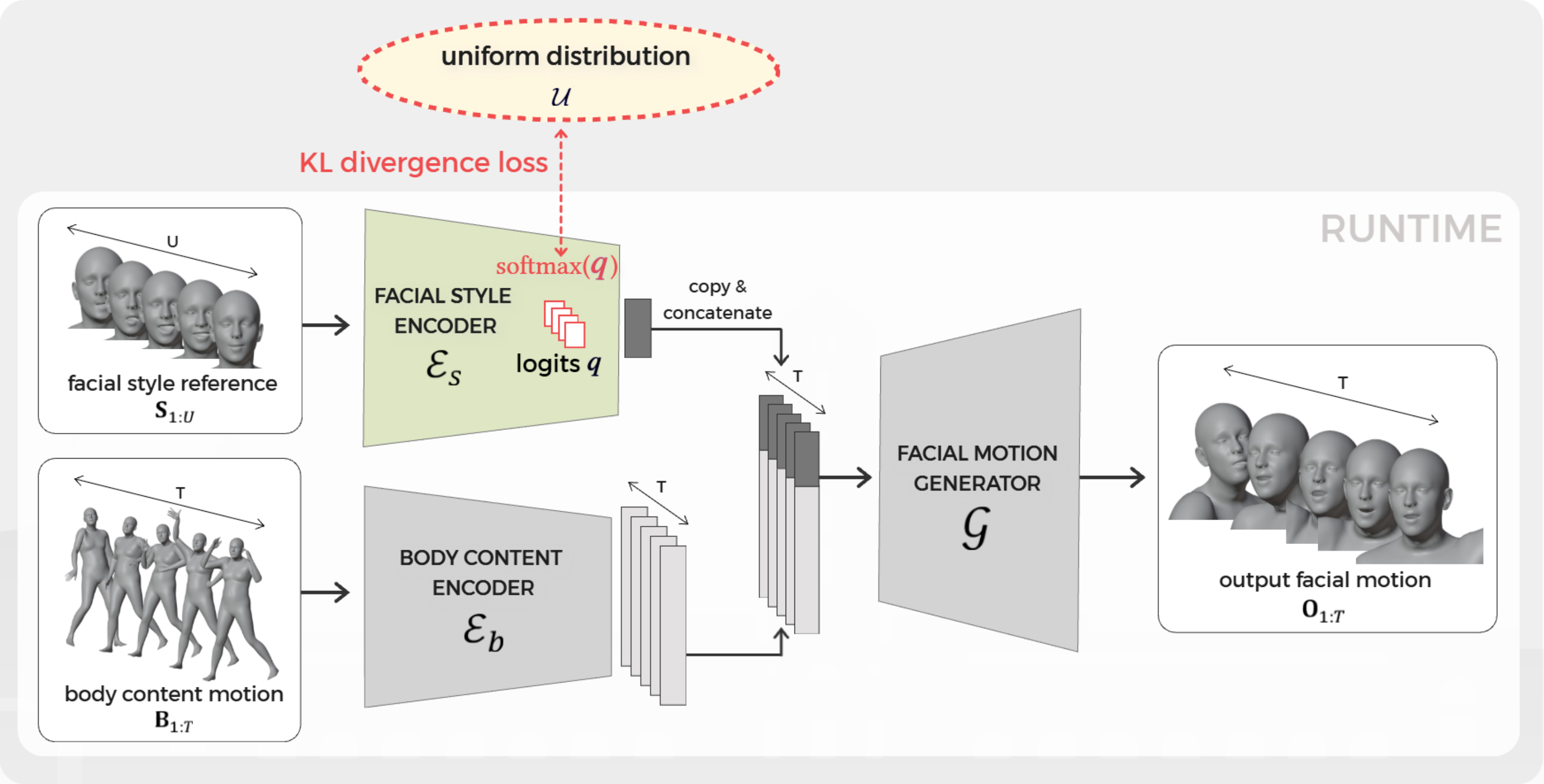
Training

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \underline{\mathcal{L}_{\text{KL}}} + \lambda_4 \mathcal{L}_{\text{consi}} + \lambda_5 \mathcal{L}_{\text{cross}}$$

- KL divergence Loss

  Minimizes KL divergence between **each categorical distribution** and a **uniform distribution.**

  Encourages balanced use of all categories



Training

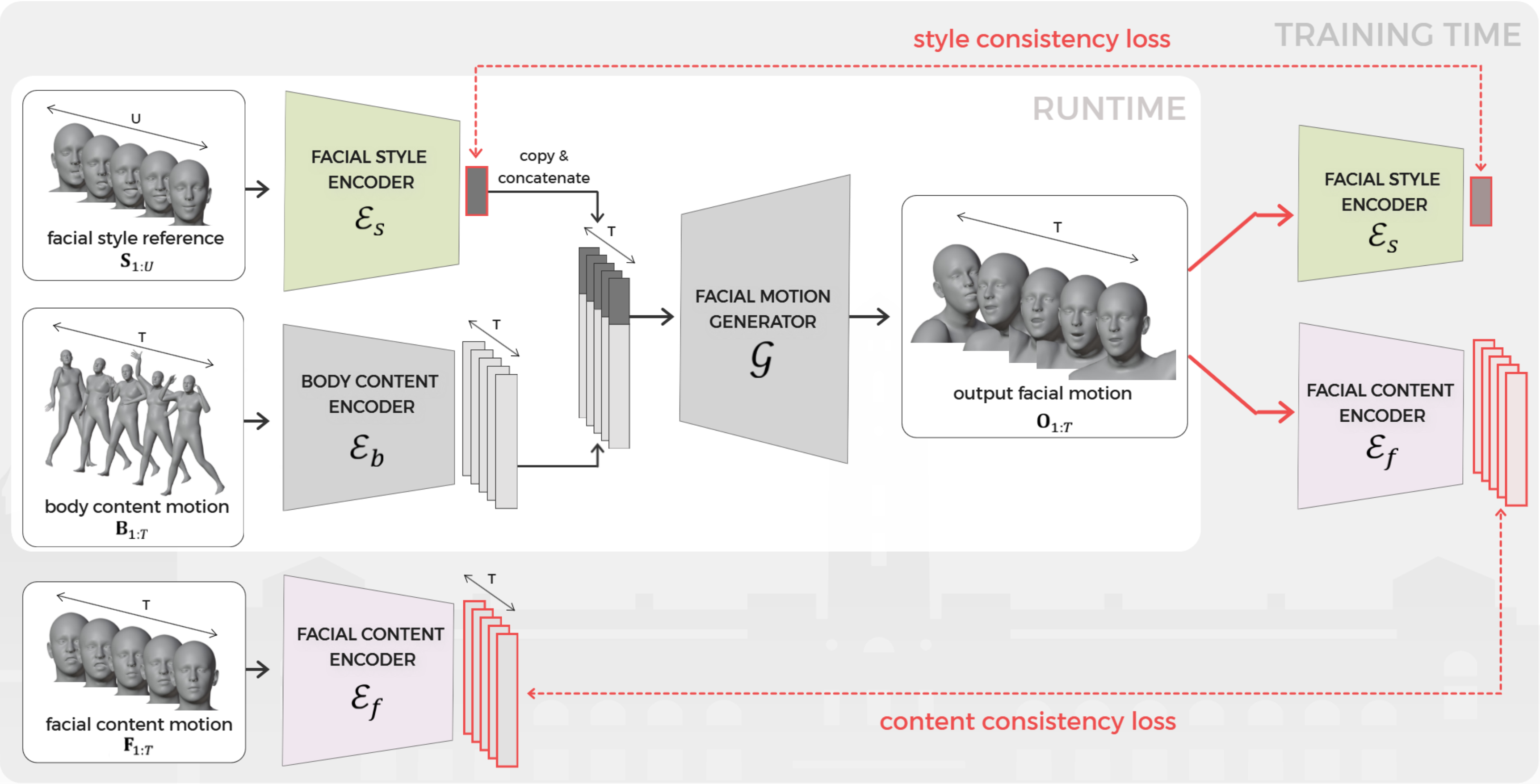
$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{KL}} + \lambda_4 \underline{\mathcal{L}_{\text{consi}}} + \lambda_5 \mathcal{L}_{\text{cross}}$$

- **Consistency Loss**

  Consist of **Style** and **Content** terms

  Compare **re-encoded output** and **original embeddings**

  Ensure generated motion reflects original content and style



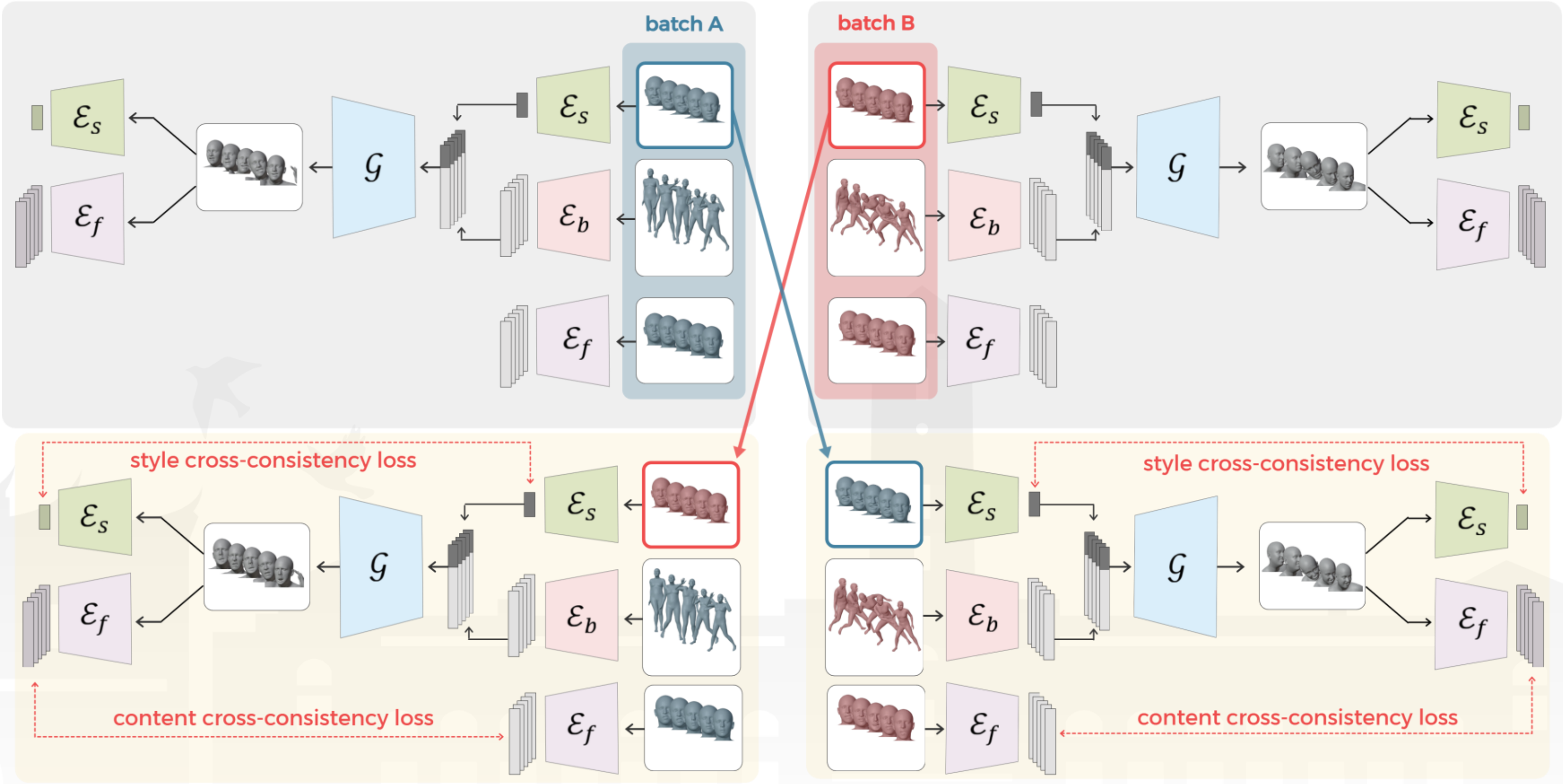
Training

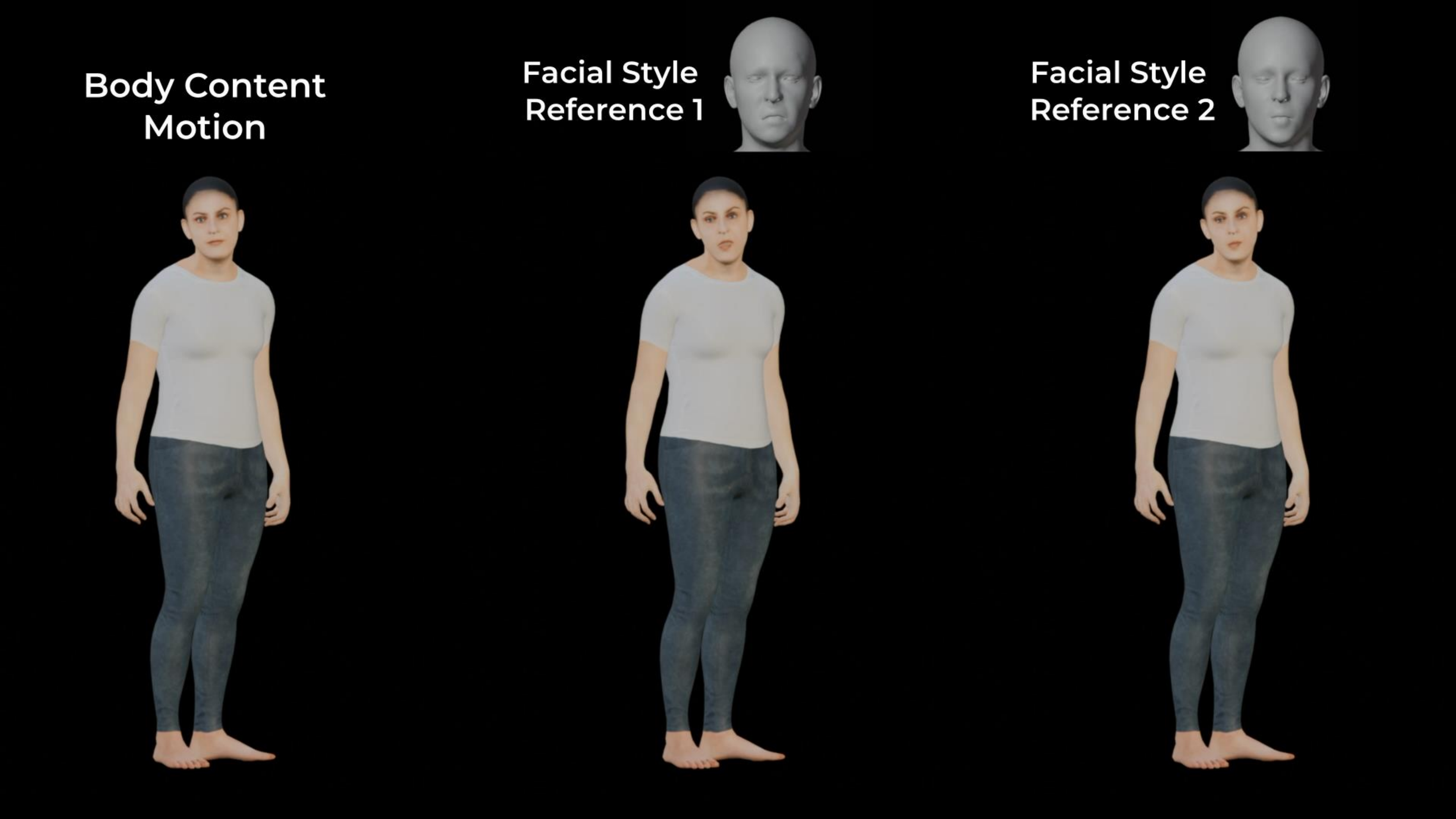
$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{KL}} + \lambda_4 \mathcal{L}_{\text{consi}} + \lambda_5 \underline{\mathcal{L}_{\text{cross}}}$$

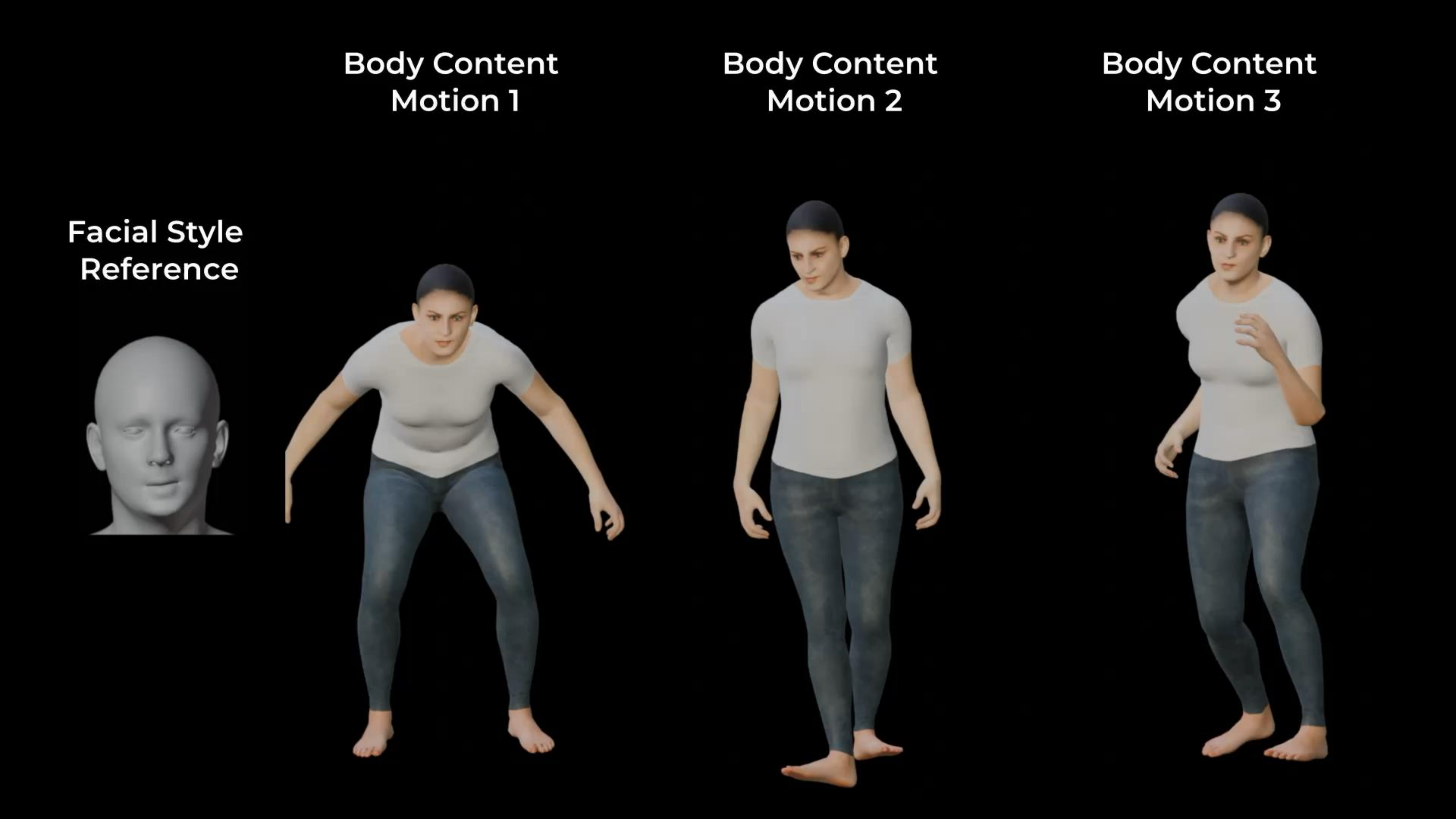
- Cross Consistency Loss

  Same structure as consistency loss, but uses content and style **from different batches**

  Ensures clear **content–style disentanglement**

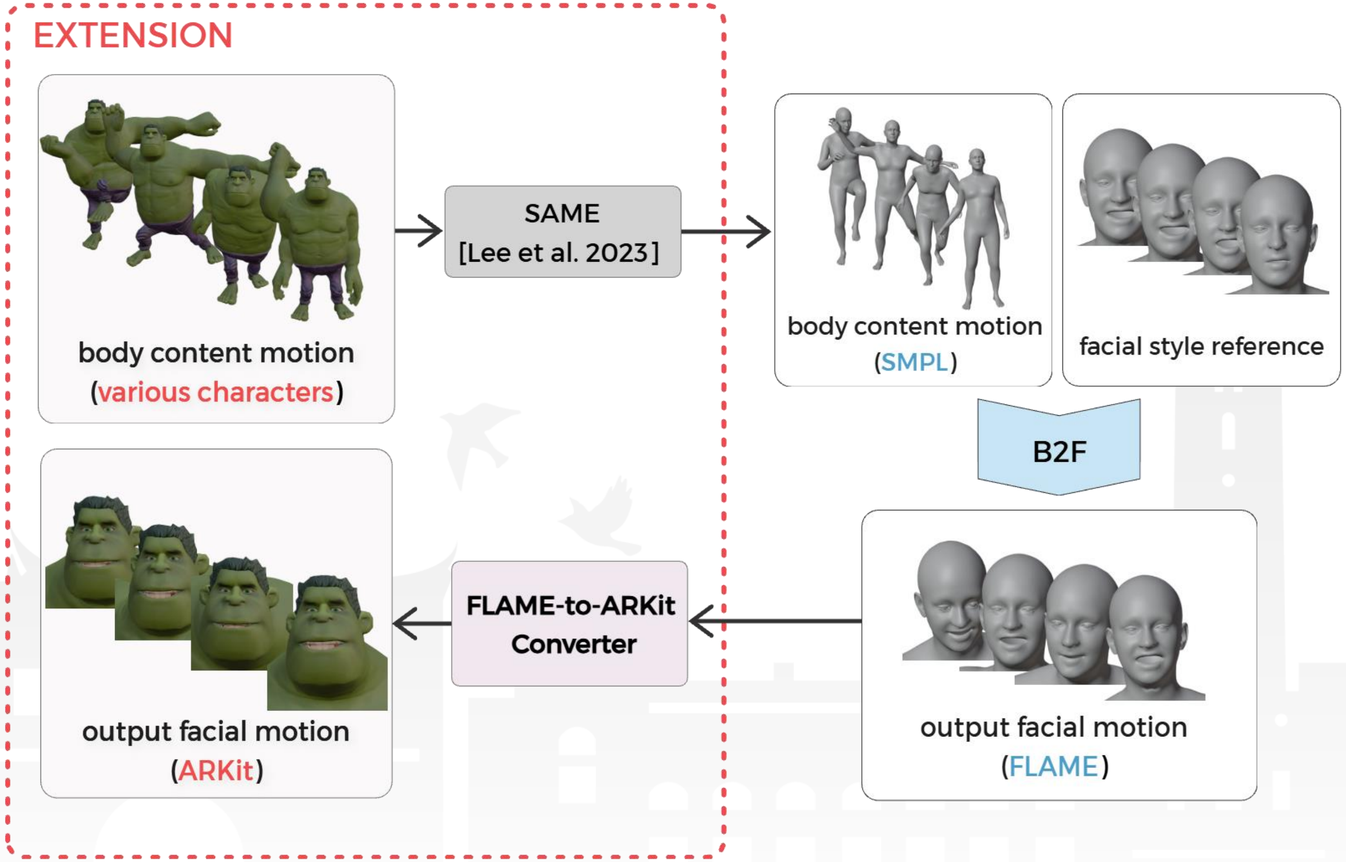



Body Content Motion
Facial Style Reference 1
Facial Style Reference 2


Facial Style Reference
Body Content Motion 1
Body Content Motion 2
Body Content Motion 3

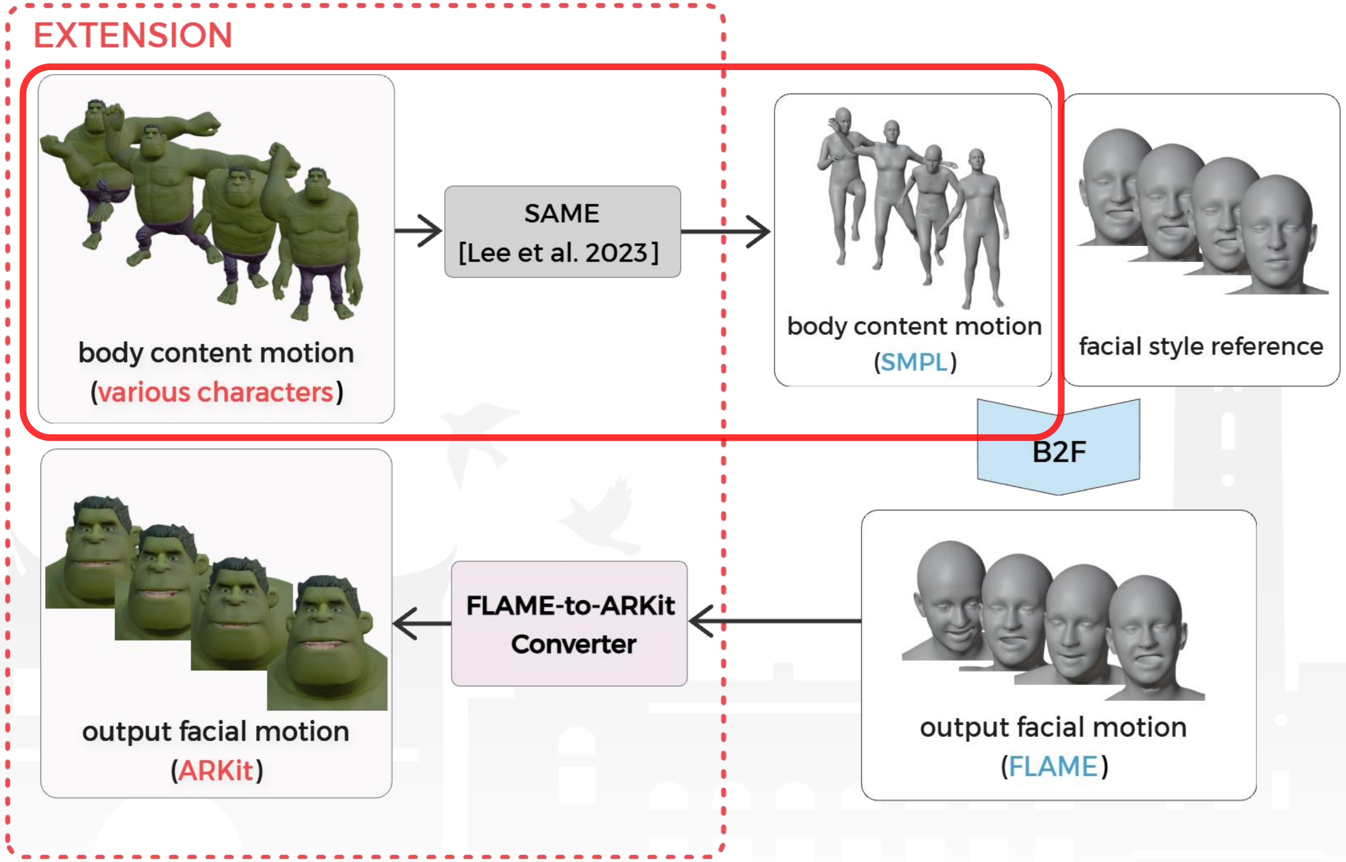
Extension

- We **extend** B2F for broader use, enabling it to work regardless of **skeleton structure** or **facial motion format.**



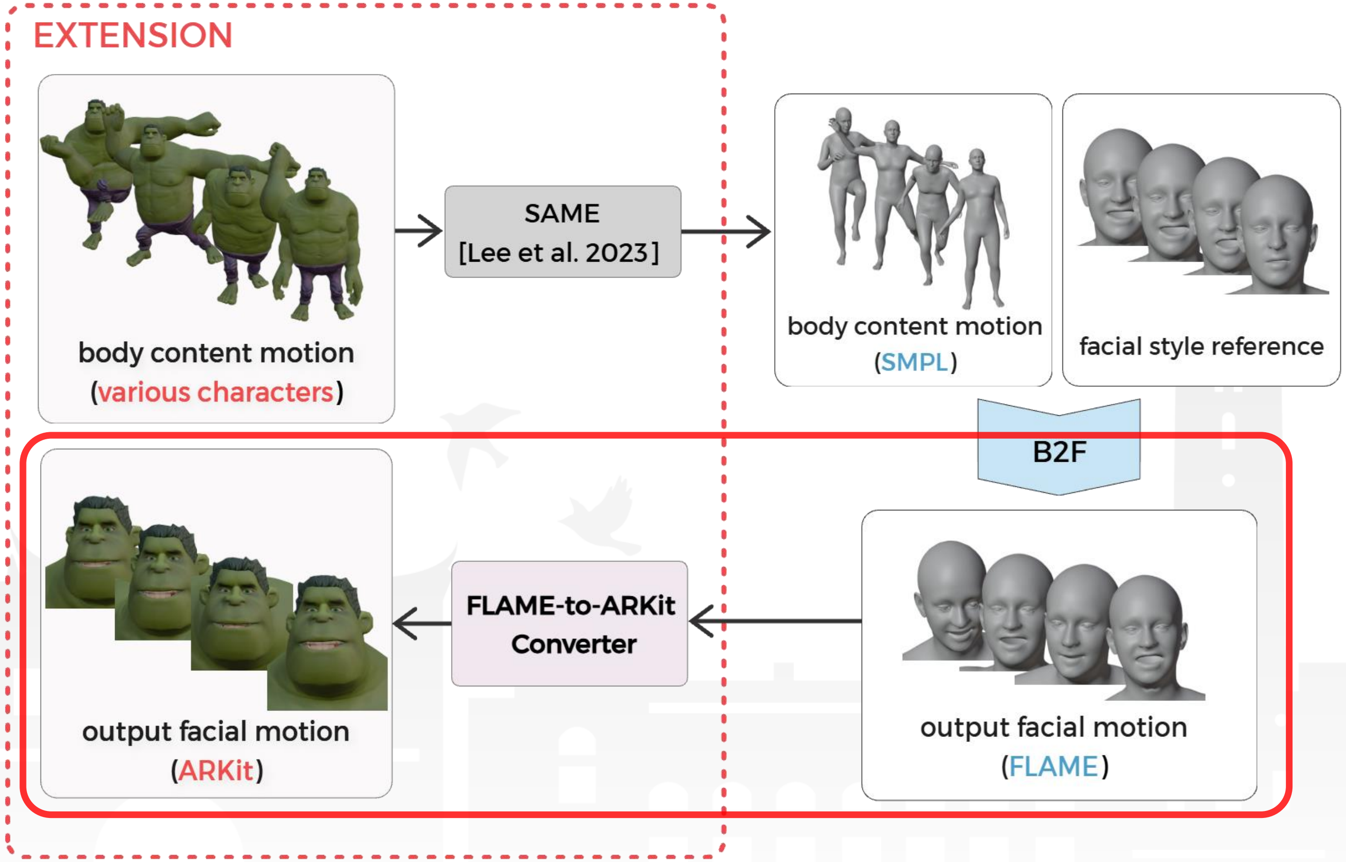
Extension

- Motion Retargeting for Input: Use SAME [LKP*23] to map diverse skeleton motions into a unified embedding for B2F input.



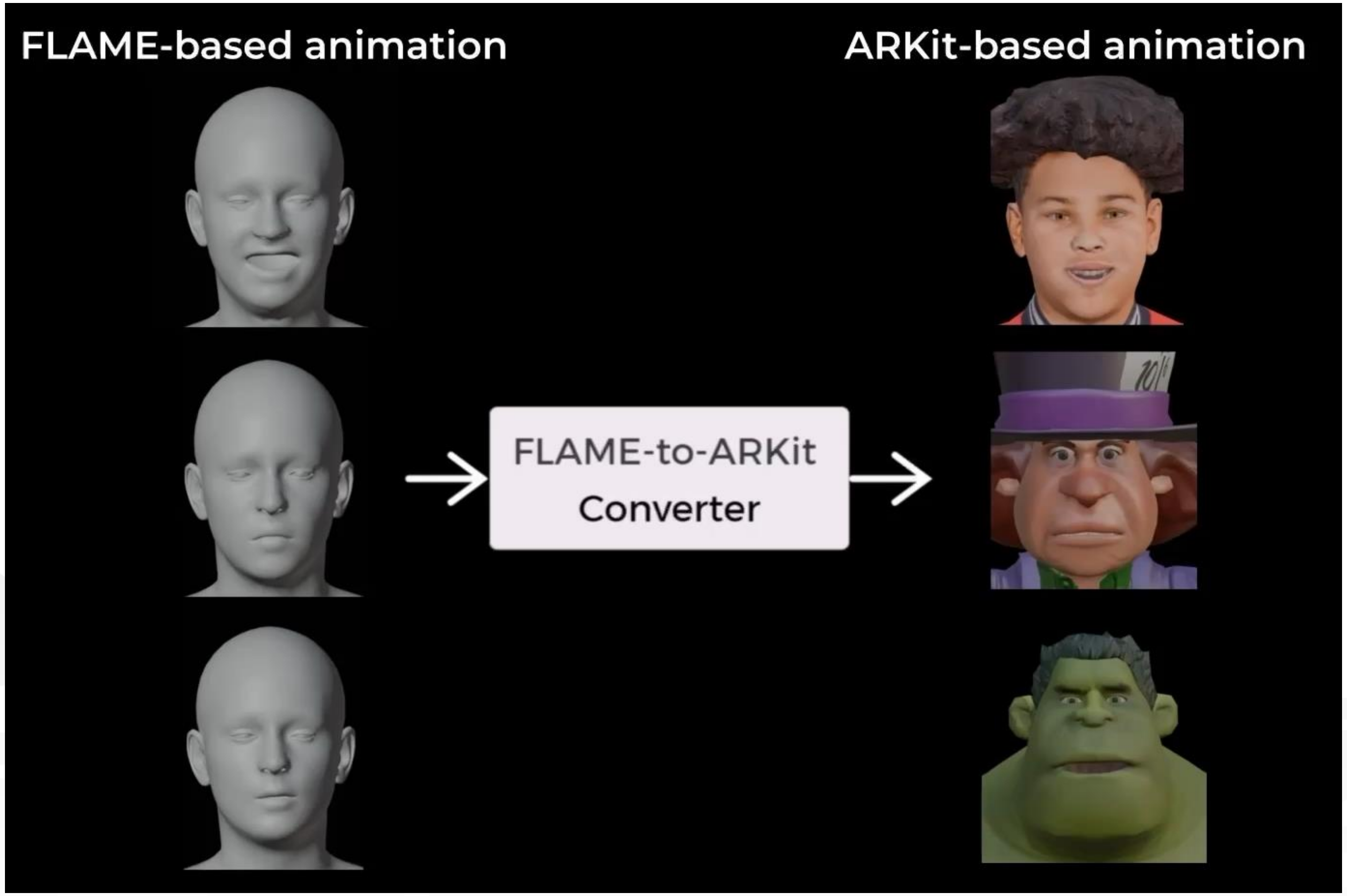
Extension

- FLAME-to-ARKit Converter for Output: Converts FLAME outputs to ARKit blendshapes for practical deployment.



Extension

- FLAME-to-ARKit Converter for Output: Converts FLAME outputs to ARKit blendshapes for practical deployment.



Body Content Motion

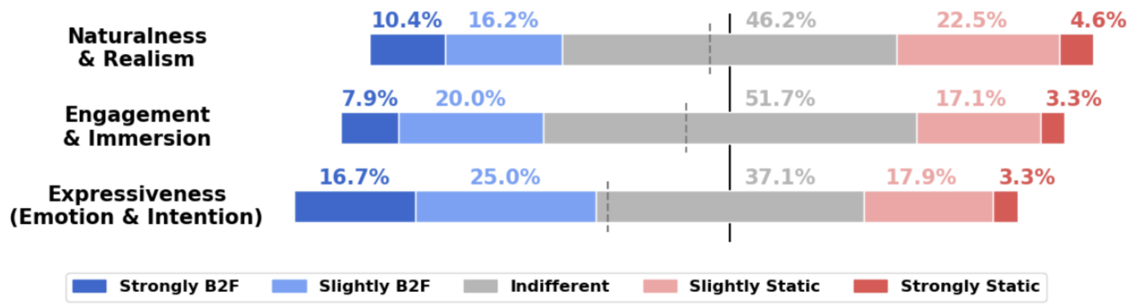
Facial Style Reference

User Studies

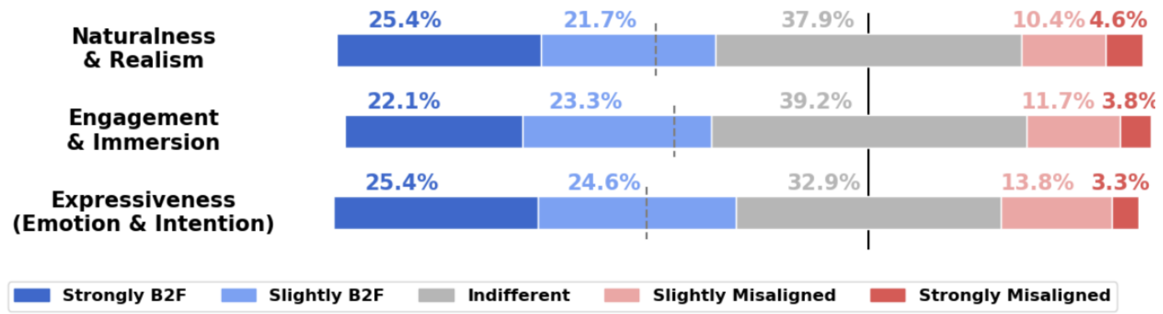
2025 TAIPEI

Two **user studies** evaluated how facial motion affects perception in terms of **naturalness**, **engagement**, and **expressiveness**.

User Study 1 **(B2F vs. Static Face)**



User Study 2 **(B2F vs. Misaligned Face)**



→ The experiments confirmed that **B2F improves perceptual quality** in both cases, especially showing a clearer advantage over the misaligned face condition.

Quantitative Evaluation

- We conducted **a quantitative evaluation** using several **ablation** and **baseline models** to assess the impact of each component and input setting

Quantitative Evaluation

- We conducted **a quantitative evaluation** using several **ablation** and **baseline models** to assess the impact of each component and input setting

- Ablation Models

  w/o $\mathcal{L}_{\textbf{align}}$
  w/o $\mathcal{L}_{\textbf{cross}}$
  w/o $\mathcal{L}_{\textbf{consi}} \& \mathcal{L}_{\textbf{cross}}$

- Baseline Models

  w/ SMPL Pose Input : Uses **SMPL body pose** as input instead of body content features.
  w/ SAME Pose Input : Uses skeleton-agnostic pose representation (**SAME**) as input.
  B2F-VAE : Uses a **fully continuous latent space** for style encoding.
  B2F-VQVAE : Uses a **discretized codebook-based** style representation.

Quantitative Evaluation

- We conducted **a quantitative evaluation** using several **ablation** and **baseline models** to assess the impact of each component and input setting

- Ablation Models

  w/o $\mathcal{L}_{\text{align}}$
  w/o $\mathcal{L}_{\text{cross}}$
  w/o $\mathcal{L}_{\text{consi}}$&$\mathcal{L}_{\text{cross}}$

- Metric

(1) ℓ_2 Error

the average l2 error between predicted and ground-truth blendshape weights to assess reconstruction accuracy

(2) Std. Dev. Difference

the average absolute difference in standard deviation across each blendshape dimension over time to assess temporal variation

- Baseline Models

w/ SMPL Pose Input : Uses **SMPL body pose** as input instead of body content features.
w/ SAME Pose Input : Uses skeleton-agnostic pose representation (**SAME**) as input.
B2F-VAE : Uses a **fully continuous latent space** for style encoding.
B2F-VQVAE : Uses a **discretized codebook-based** style representation.

Quantitative Evaluation

2025 TAIPEI

- We conducted **a quantitative evaluation** using several **ablation** and **baseline models** to assess the impact of each component and input setting

Lowest reconstruction error →

Model	ℓ_2 Error ↓	Std. Dev. Difference ↓
B2F (ours)	0.556	0.309
w/o $\mathcal{L}_{\text{align}}$	0.593	0.331
w/o $\mathcal{L}_{\text{cross}}$	0.591	0.326
w/o $\mathcal{L}_{\text{consi}}$&$\mathcal{L}_{\text{cross}}$	0.565	0.297
w/ SMPL Pose Input	0.618	0.381
w/ SAME Pose Input	0.673	0.564
B2F-VAE	0.567	0.321
B2F-VQVAE	0.570	0.292

Quantitative Evaluation

- We conducted **a quantitative evaluation** using several **ablation** and **baseline models** to assess the impact of each component and input setting

Model	ℓ_2 Error ↓	Std. Dev. Difference ↓
B2F (ours)	0.556	0.309
w/o $\mathcal{L}_{align}$	0.593	0.331
w/o $\mathcal{L}_{cross}$	0.591	0.326
w/o $\mathcal{L}_{consi}$&$\mathcal{L}_{cross}$	0.565	0.297
w/ SMPL Pose Input	0.618	0.381
w/ SAME Pose Input	0.673	0.564
B2F-VAE	0.567	0.321
B2F-VQVAE	0.570	0.292

lower temporal deviation error than ours

Quantitative Evaluation

- We conducted **a quantitative evaluation** using several **ablation** and **baseline models** to assess the impact of each component and input setting

Model	ℓ_2 Error ↓	Std. Dev. Difference ↓
B2F (ours)	0.556	0.309
w/o $\mathcal{L}_{\text{align}}$	0.593	0.331
w/o $\mathcal{L}_{\text{cross}}$	0.591	0.326
w/o $\mathcal{L}_{\text{consi}}$&$\mathcal{L}_{\text{cross}}$	0.565	0.297
w/ SMPL Pose Input	0.618	0.381
w/ SAME Pose Input	0.673	0.564
B2F-VAE	0.567	0.321
B2F-VQVAE	0.570	0.292

Less expressive. Better score only for this evaluation setting (when style and content come from the same clips)

Numerically good, but shows unnatural expressions due to rigid style discretization.

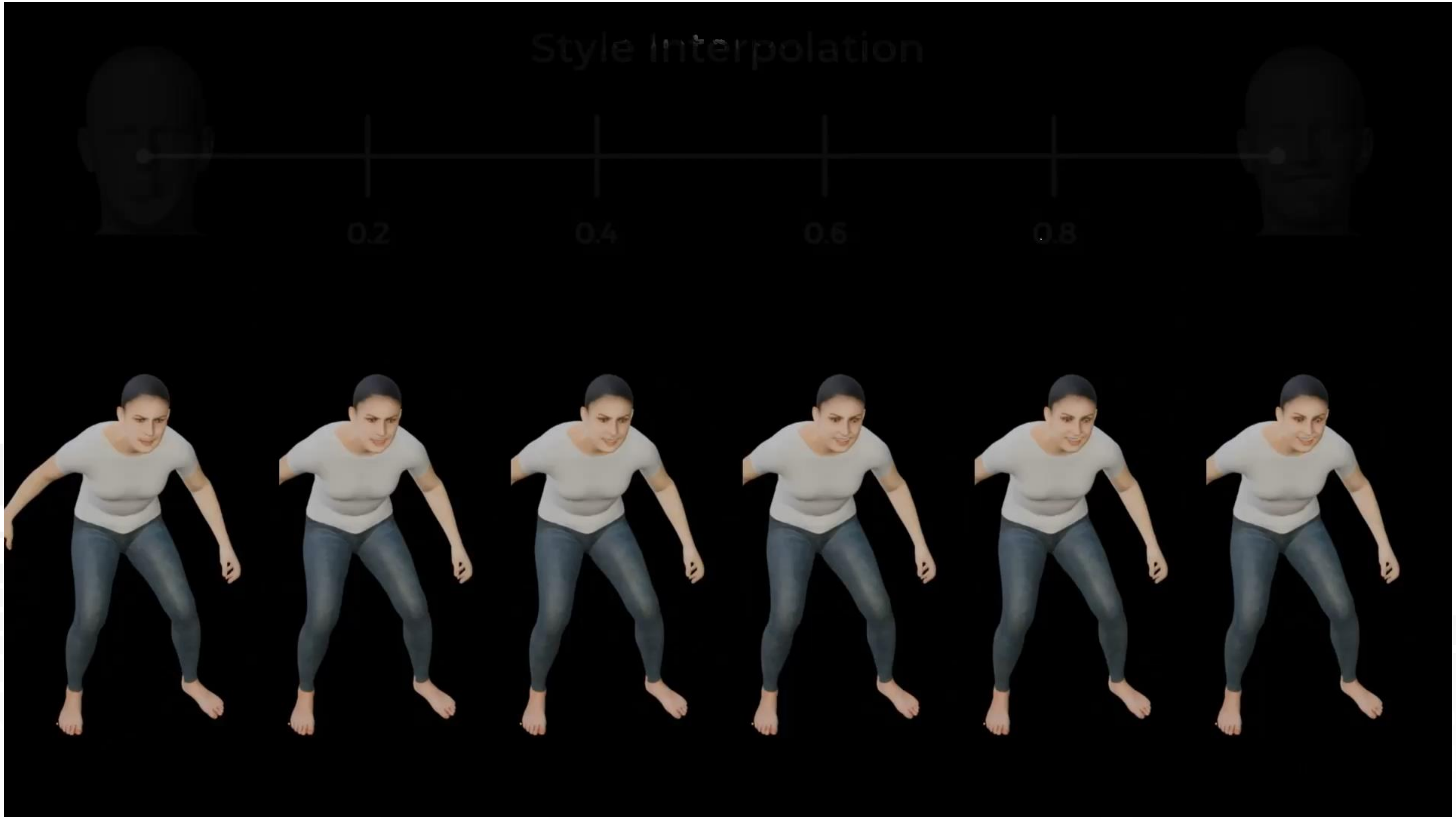
B2F: End-to-End Body-to-Face Motion Generation with Style Reference

Bokyung Jang, Eunho Jung, Yoonsang Lee*

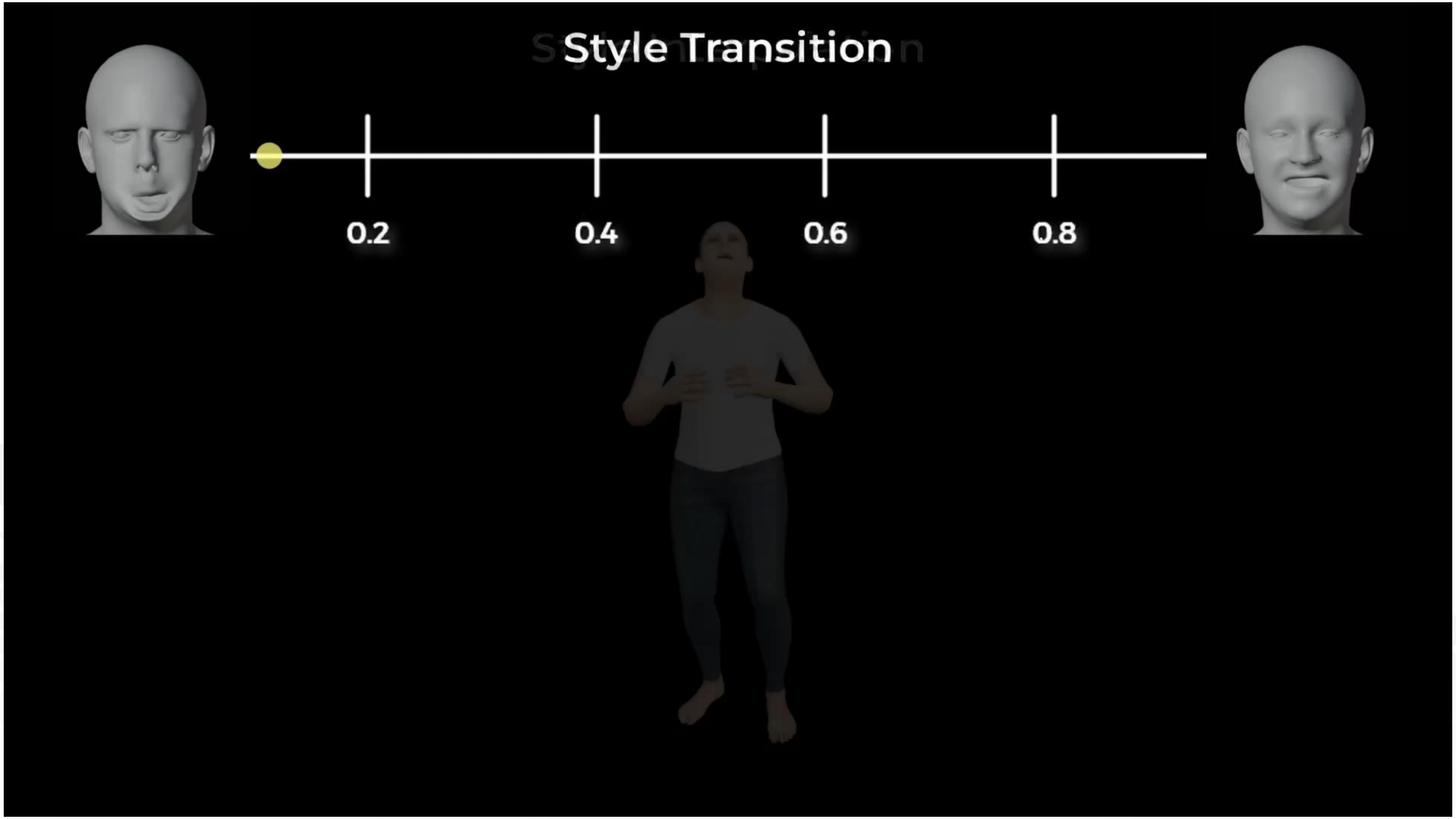
Hanyang University

2025 TAIPEI

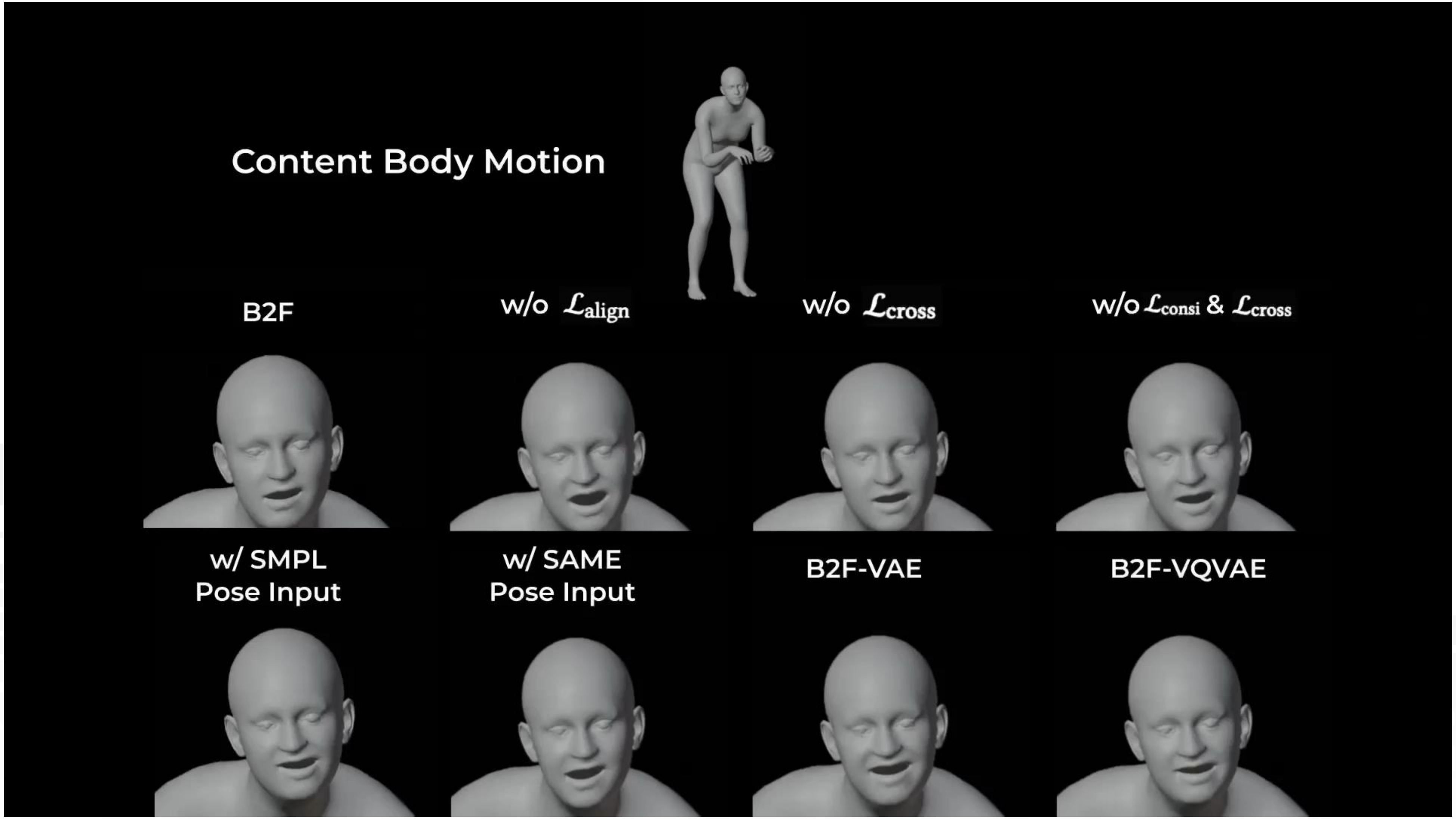
Style Interpolation



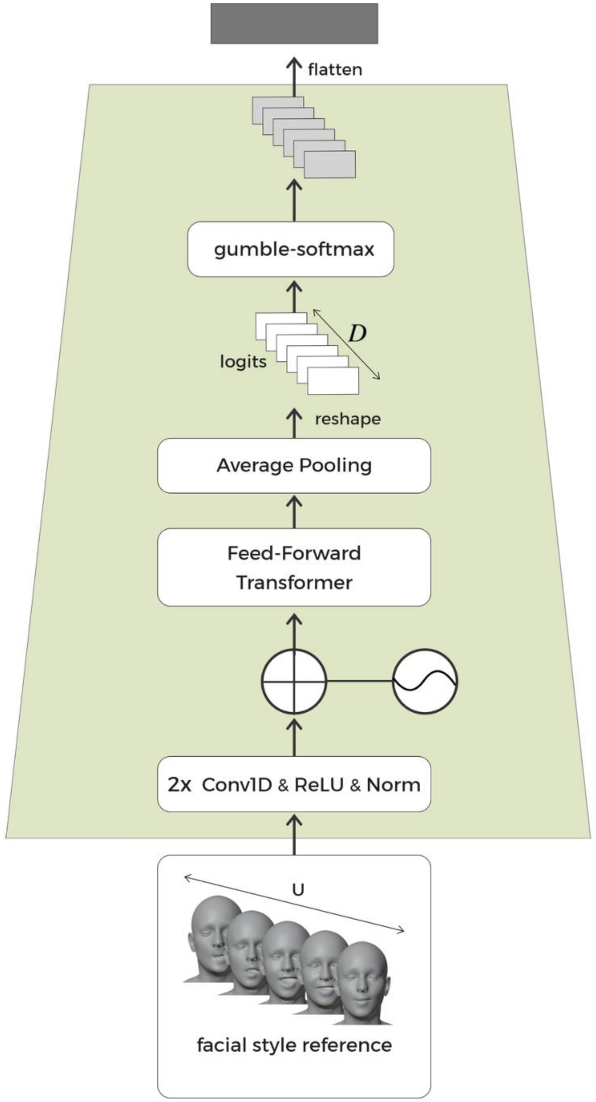
Style Transition



Ablation Study



Gumble-Softmax



- Uses **Gumbel-Softmax** to sample soft categorical vectors from multiple distributions.
- Each sampled vector represents a different **style region**, combined into one style embedding.
- Overcomes VAE’s blurry styles and VQ-VAE’s unstable training, keeping styles **clear, diverse, and stable**.