A. Supplementary Material

A.1. Weighting Factors in the Reconstruction Loss of VAE Variants

Name	Value	
	BaseVAE / *StructVAE*	*MultiObjVAE* / *FactorVAE*
$w^{\text{pos}}_{\text{T-A}}$	14.5	5
$w^{\text{pos}}_{\text{RH}}$	14.5	10
$w^{\text{pos}}_{\text{LH}}$	14.5	12
$w^{\text{pos}}_{\text{F}}$	14.5	4
$w^{\text{pos}}_{\text{M}}$	0	8
$w^{\text{vel}}_{\text{T-A}}$	0	3
$w^{\text{vel}}_{\text{F}}$	0	6
$w^{\text{vel}}_{\text{M}}$	0	10

Table 1. Weights used in $\mathcal{L}_{\text{recon}}$ depending on the VAE variant (Equations 2 and 3).

A.2. Qualitative Comparison of Reconstructed and Generated Sign Pose Sequences

Figure 1 presents: 1) reconstructed poses from a ground-truth sign pose sequence of the PHOENIX14T dataset using *MultiObjVAE* and *FactorVAE*, and 2) the generated sequences from the corresponding input sentence[1] using the latent diffusion model trained on each VAE latent space. Despite visually similar VAE reconstructions for both variants (notably with correct hands and arms motions), we observe meaningful differences between generated sequences.

[1] *"Und nun die Wettervorhersage für morgen, Donnerstag, den sechsundzwanzigsten November."* meaning *"And now the weather forecast for tomorrow, Thursday, the twenty-sixth of November."*

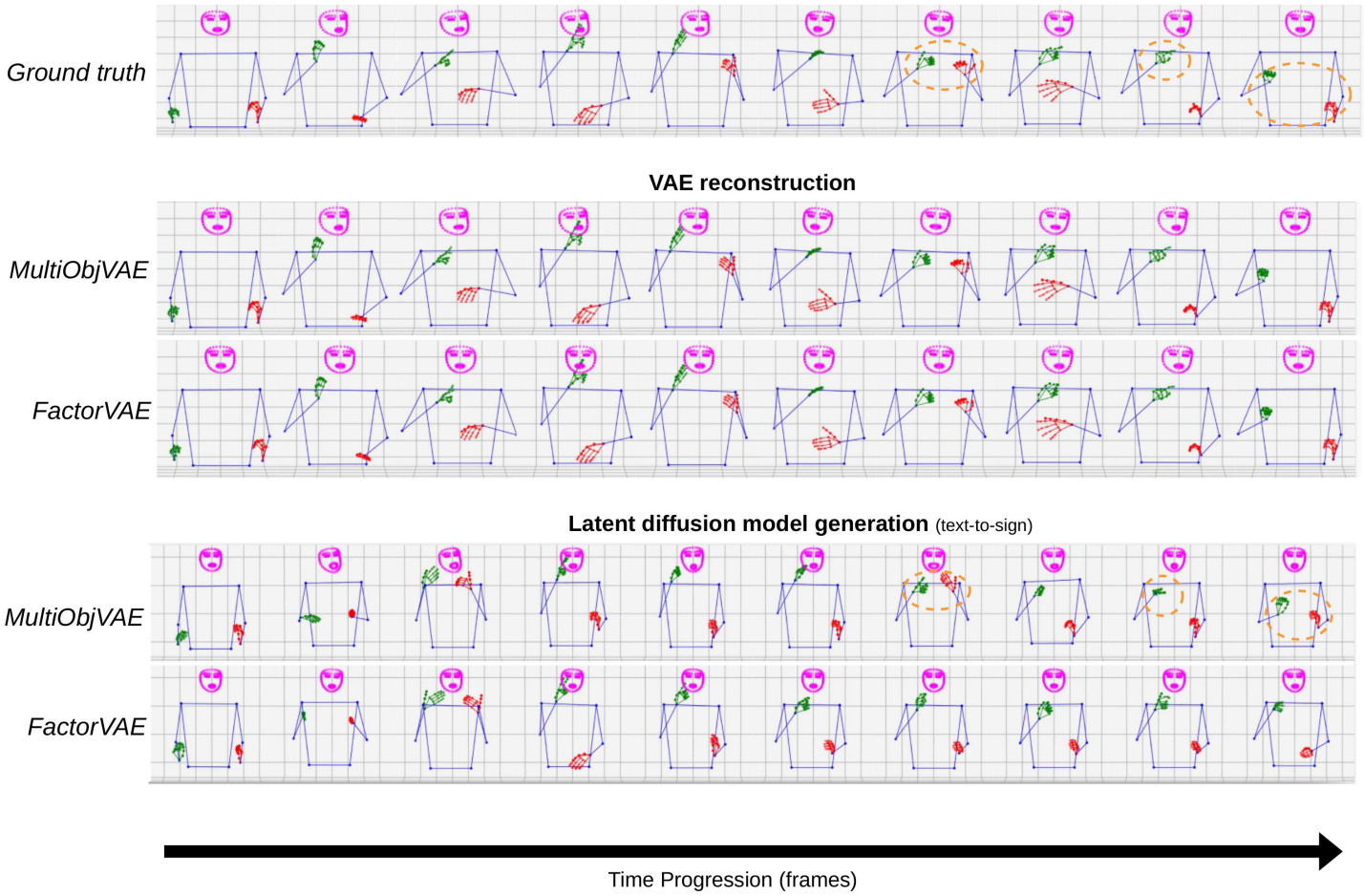

Figure 1. Qualitative comparison of sign pose sequences. Top: ground-truth poses and VAE reconstructions produced by *MultiObjVAE* and *FactorVAE*. Bottom: text-to-sign sequences generated by latent diffusion models trained on the corresponding latent spaces. Orange doted circles indicate hand configurations correctly generated in the *MultiObjVAE* latent space but not in the *FactorVAE* case.
Sequence ID in the PHOENIX14T dataset: `25November_2009_Wednesday_tagesschau-7666`.