Advanced Sign Language Video Generation with Compressed and Quantized Multi-Condition Tokenization

Cong Wang ; Zexuan Deng ; Zhiwei Jiang ; Yafeng Yin ; Fei Shen ; Zifeng Cheng ; Shiping Ge ; Shiwei Gan ; Qing Gu

Introduction

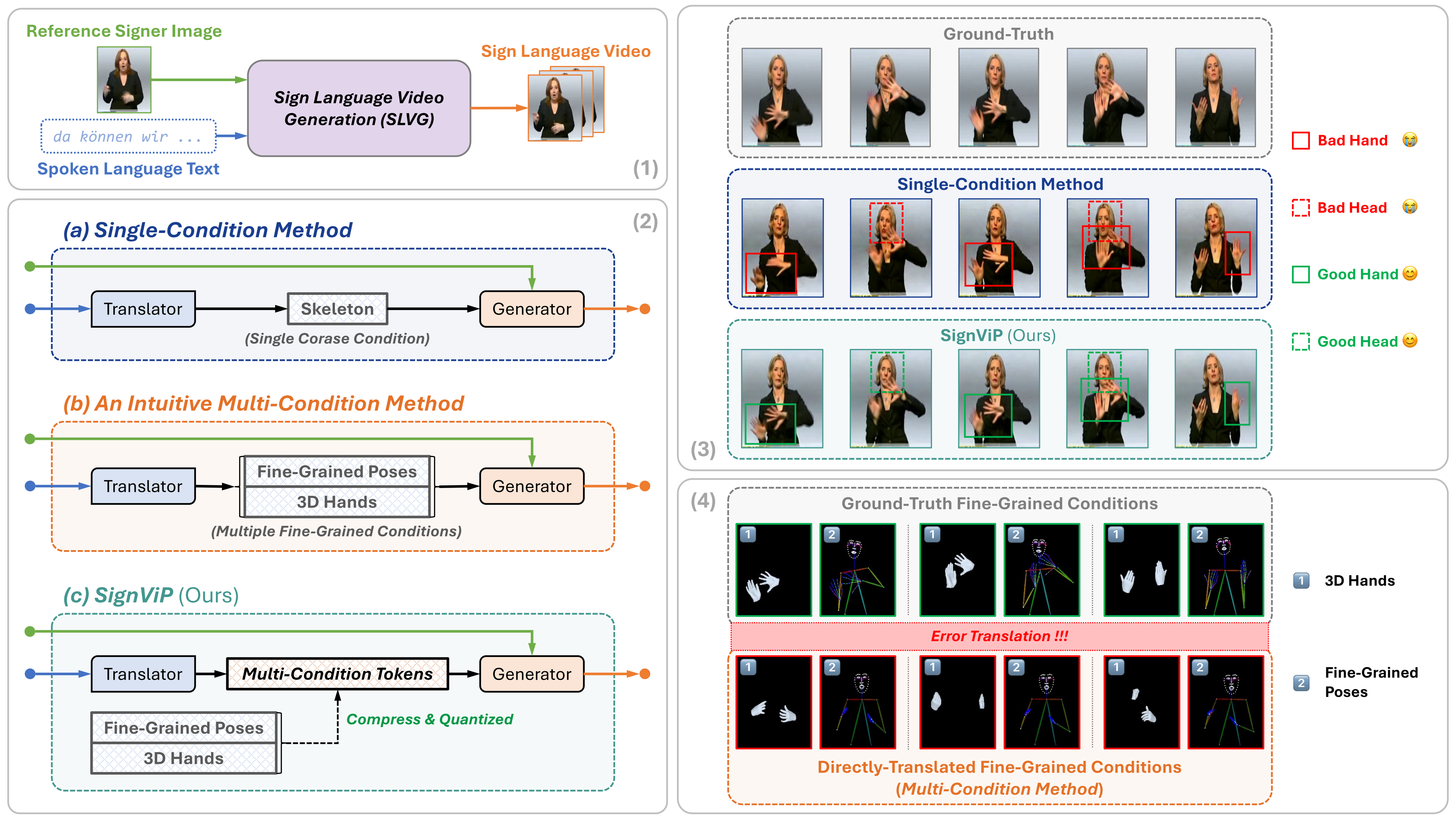
Fig. (1) Sign Language Video Generation (SLVG):
generate sign language video from signer image and spoken language text.

Fig. (2a) Naïve Solution #1:
Skeleton (single coarse condition) to bridge Translator and Generator. **(poor performance, Fig. (3))**

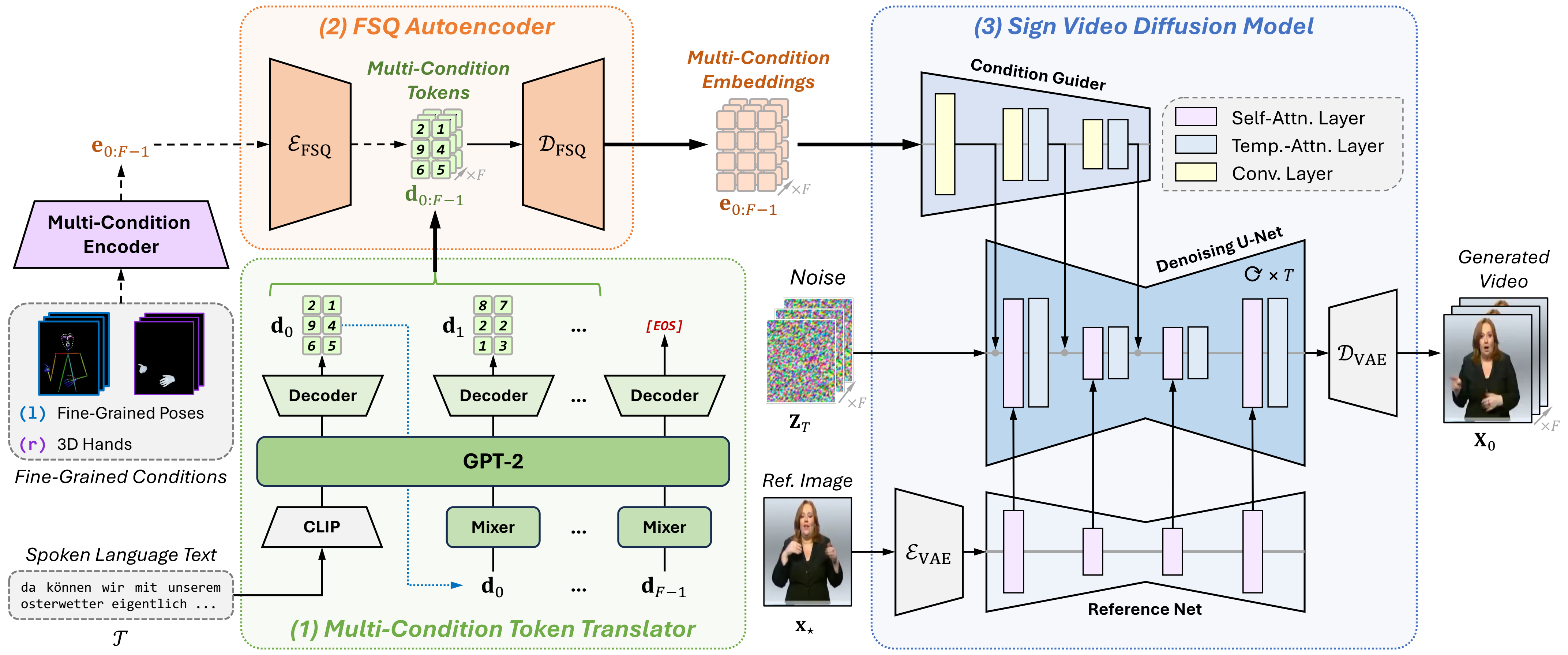
Fig. (2b) Naïve Solution #2:
~~Skeleton~~ → Multiple Fine-Grained Conditions. **(difficult to translate, Fig. (4))**

How to overcome the challenges in
the translation of multiple fine-grained conditions?

SignViP adopts ***a discrete tokenization paradigm*** to
integrate and represent ***multiple fine-grained conditions***.
(Fig.2(c))



Method



the construction of ***a discrete multi-condition token space***

Sign Language Video   Spoken Language Text

$$p_\theta(\mathbf{X}|\mathbf{x}_\star,\mathcal{T}) = p_{\theta_{\text{gen}}}(\mathbf{X}|\mathbf{x}_\star,\mathbf{e}_{0:F-1}) \cdot p_{\theta_{\text{AE}}}(\mathbf{e}_{0:F-1}|\mathbf{d}_{0:F-1}) \cdot p_{\theta_{\text{tran}}}(\mathbf{d}_{0:F-1}|\mathcal{T})$$

Signer Image   Continuous Multi-Condition Embeddings   Discrete Multi-Condition Tokens

Experiments

Table 1: Comparison of **video back-translation** performance.

	RWTH-2014T						How2Sign					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	COMET	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	COMET
Ground-Truth	33.06	20.81	15.00	11.90	34.27	0.6157	20.37	13.11	9.78	7.53	21.43	0.5882
MoMP [59] + ControlNet [88]	19.12	8.95	5.33	3.61	21.54	0.5033	12.53	5.59	3.48	2.31	13.72	0.5122
MoMP [59] + AnimateAnyone [29]	20.05	8.79	5.24	3.72	21.68	0.5091	13.65	5.82	3.39	2.25	14.15	0.5208
SignGAN [61]	17.41	7.93	4.67	3.16	19.64	0.4977	10.66	4.62	2.92	1.97	11.76	0.5104
w/ AnimateAnyone [29]	18.29	7.75	4.59	3.23	19.70	0.4928	11.82	4.85	2.83	1.92	12.12	0.5135
SignGen [47]	13.28	3.05	1.13	0.51	16.13	0.4086	8.21	1.91	0.64	0.41	9.54	0.4127
SignViP (Ours)	**26.72**	**15.65**	**11.14**	**8.65**	**28.85**	**0.5608**	**16.21**	**9.36**	**6.28**	**5.04**	**16.99**	**0.5524**

Table 2: Comparison of **pose back-translation** performance.

	RWTH-2014T						How2Sign					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	COMET	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	COMET
Ground-Truth	30.99	18.36	12.83	9.87	31.02	0.5978	24.56	14.96	10.31	7.91	24.88	0.6250
ProTran [58]	17.96	8.99	5.64	4.07	20.97	0.5091	14.57	7.47	4.59	3.42	17.32	0.5549
Adversarial [56]	17.70	8.96	5.72	4.18	21.15	0.5127	14.76	7.15	4.66	3.48	17.84	0.5618
MDN [60]	18.06	9.30	6.06	4.52	21.44	0.5251	14.94	7.54	5.10	3.67	18.21	0.5685
MoMP [59]	20.55	**10.98**	**7.02**	**5.14**	**23.75**	**0.5466**	16.57	**8.47**	5.38	4.16	**19.38**	**0.5802**
SignGAN [61]	12.14	6.10	3.88	2.85	14.79	0.5123	10.86	5.27	3.30	2.62	13.19	0.5673
w/ AnimateAnyone [29]	12.36	6.23	4.01	2.97	14.93	0.5231	10.97	5.36	3.39	2.67	13.36	0.5315
SignGen [47]	10.42	2.42	0.89	0.38	12.68	0.4324	8.67	1.79	2.41	1.36	8.69	0.4413
SignViP (Ours)	**21.94**	10.06	6.32	4.61	22.67	0.5347	**17.35**	8.28	**5.41**	**4.42**	18.23	0.5738

Table 3: Comparison of video quality.

	RWTH-2014T				How2Sign			
	FID ↓	CLIP-FID ↓	FVD ↓	IDS ↑	FID ↓	CLIP-FID ↓	FVD ↓	IDS ↑
SignGAN [61]	547.90	167.70	1431.38	0.463	667.44	210.11	2766.97	0.538
w/ AnimateAnyone [29]	595.99	161.97	1330.54	0.462	679.41	215.05	2484.39	0.533
SignGen [47]	644.06	184.66	1715.32	0.515	815.69	186.32	3538.49	0.539
SignViP (Ours)	**508.91**	**154.10**	**1025.45**	**0.571**	**575.67**	**109.61**	**2207.67**	**0.624**

Table 4: Generative capability comparison of video diffusion models.

	RWTH-2014T					How2Sign				
	FVD ↓	SSIM ↑	PSNR ↑	LPIPS ↓	Hand SSIM ↑	FVD ↓	SSIM ↑	PSNR ↑	LPIPS ↓	Hand SSIM ↑
ControlNet [88]	556.63	0.784	19.50	0.137	0.483	427.22	0.826	21.32	0.116	0.657
AnimateAnyone [29]	365.42	0.794	20.06	0.121	0.505	293.18	0.821	21.54	0.103	0.663
Sign Video Diffusion Model (Ours)	**275.22**	**0.829**	**22.91**	**0.089**	**0.614**	**210.63**	**0.855**	**23.11**	**0.074**	**0.752**

Qualitative comparison

Identity Generalization