SIGNDIT: HAND-FAITHFUL POSE-GUIDED VIDEO GENERATION FOR SIGN LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Pose-guided video generation can produce plausible human motion while corrupting the small hand articulations that determine meaning in sign language. This is not merely a resolution problem: conditioning must allocate geometric information across sources, adapt their reliability over space and time, and preserve discrete linguistic contrasts. We introduce SIGNDIT, a two-stage framework for these decisions. Stage 1 assigns image-plane structure to a 2D keypoint path and projection-ambiguous surface orientation, relative depth, and occlusion order to a second path, then modulates both with a part-conditioned continuous reliability field. Stage 2, Sign-DPO, changes one phonological factor at a time to construct counterfactual preference pairs and localizes credit to the affected region and frames. In self-reenactment and cross-identity evaluation on How2Sign, CSL-Daily, and Phoenix, SIGNDIT is best or tied best on 32 of 36 fidelity measures, reduces cross-identity FID by 25.1–41.2%, comes within 1.83 BLEU-4 points of real-video back-translation, and exceeds the strongest generated baseline by 3.22–3.79 points. The resulting design treats pose control as structured inference rather than feature concatenation.

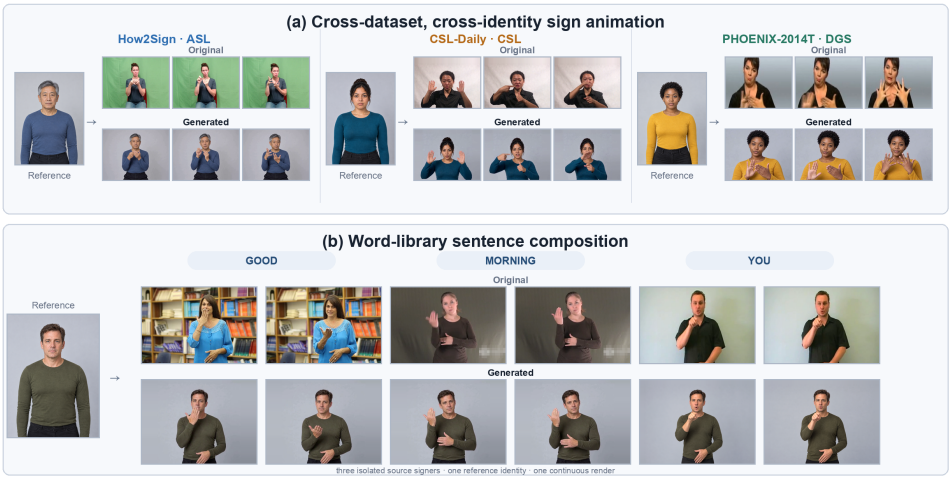

Figure 1: Capabilities of SIGNDIT. **Top:** Cross-identity sign animation on How2Sign, CSL-Daily, and Phoenix. **Bottom:** Word-library composition of GOOD MORNING YOU, rendered continuously with one reference identity.

1 INTRODUCTION

Generating a signer from a reference image and driving motion is unusually unforgiving: a bent finger or reversed palm that is a local defect in generic animation can change a sign's lexical identity. Photorealism and temporal smoothness are therefore insufficient without preserving contrasts in handshape, location, movement, orientation, and non-manual markers (Brentari, 1999; Sandler & Lillo-Martin, 2006).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Figure 1 previews cross-identity animation on ASL, CSL, and DGS benchmarks and identity-consistent composition from isolated lexical motions.

Modern video diffusion models condition image-to-video backbones on 2D keypoints, rendered bodies, or both (Hu, 2024; Xu et al., 2024; Zhu et al., 2024). Yet a hand occupies only a few latent tokens after spatial compression despite its many articulated degrees of freedom (Cheng et al., 2025); duplicating projected position and silhouette across paths spends that capacity on repeated information. Moreover, reconstruction losses do not distinguish a correct sign from a visually plausible minimal pair (Ventura et al., 2020; Saunders et al., 2022).

We formulate pose conditioning as three inference decisions: *what* each source provides, *when and where* it is reliable, and *which* deviations alter meaning. In Stage 1 of SIGNDIT (Figure 2), a 2D path carries image-plane keypoints while a projection-complementary path carries relative depth, surface normals, and occlusion order from an articulated 3D fit. A part-conditioned reliability field modulates both sources over body regions and video time.

Stage 2 changes one phonological factor while holding the remaining motion fixed. Sign-DPO scores the resulting preferred and counterfactual samples under the original condition, localizes credit to the edited region and frames, and factorizes preferences over handshape, location, movement, and orientation. A training-time recognizer filters ambiguous pairs and calibrates their semantic margin; separate models evaluate the final videos.

Our contributions are:

- We separate image-plane pose from depth, orientation, and occlusion in a projection-complementary dual-path representation, compared against 2D-only, full-render, and additive multi-cue alternatives.
- A part-conditioned continuous reliability field uses estimator diagnostics to modulate both sources and improves over framewise and ungated controls.
- Sign-DPO uses one phonological intervention to define the hard negative, preference factor, and spatiotemporal credit mask, aligning optimization with minimal-pair semantics.
- Evaluation spans visual and pose fidelity, sign recognizers, minimal-pair tests, and a blinded fluent-signer study.

2 RELATED WORK

Sign-language video generation. Sign-language production is sensitive to articulation errors because handshape, location, movement, orientation, and non-manual markers carry linguistic contrast (Brentari, 1999; Sandler & Lillo-Martin, 2006; Duarte et al., 2021). Earlier systems combined text-to-pose and adversarial rendering, exposed the hand-detail limits of 2D transfer, and scaled dictionary-based co-articulation (Stoll et al., 2020; Ventura et al., 2020; Saunders et al., 2022). Diffusion methods add skeletal or multimodal sign conditions (Fang et al., 2025; Qi et al., 2024), compress fine-grained pose and 3D hands into tokens (Wang et al., 2025), or factorize and locally adapt pose transfer (He et al., 2026; Pelykh et al., 2026). SIGNDIT instead separates image-plane evidence from projection-ambiguous 3D cues and uses phonological contrast for alignment.

Sign-language understanding for evaluation. Semantic evaluation requires language-matched models rather than generic video metrics. CV-SLT aligns visual and textual representations (Zhao et al., 2024), SpaMo combines spatial and motion encoders with an LLM (Hwang et al., 2025), and Uni-Sign unifies recognition and translation under large-scale pretraining (Li et al., 2025b); related systems use multimodal experts or long-context reconstruction (Thomas et al., 2026; Han et al., 2026). Our evaluator pool requires released checkpoints and language-matched calibration.

Pose-guided human animation. Pose-guided animation combines landmarks with denser appearance or geometry: reference and pose attention (Hu, 2024; Xu et al., 2024), skeleton plus SMPL-derived maps (Zhu et al., 2024), confidence-aware guidance (Zhang et al., 2025), or motion fields from sparse keypoints (Li et al., 2025a). DiT systems minimally adapt strong video priors (Zhou et al., 2025), align skeleton and face features (Cheng et al., 2025), or learn unified 3D-consistent pose and driving latents (Yan et al., 2025; 2026). Rather than augmenting or unifying overlapping controls,

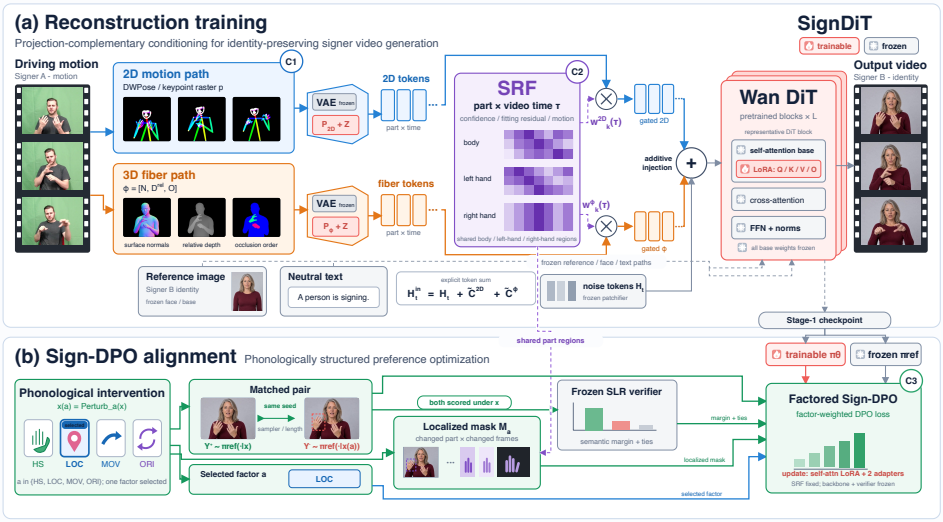

Figure 2: Overview of SIGNDIT. (a) Signer A drives the complementary 2D and fiber paths, gated by the SRF before additive injection. (b) A phonological intervention jointly defines the matched pair, selected factor, and localized mask for Sign-DPO.

our projection-fiber view assigns projected motion and complementary 3D information to distinct paths before injection.

Reliability-aware conditioning and implicit fields. Pose reliability varies across articulators and time. MimicMotion exposes confidence to the generator (Zhang et al., 2025), while SIREN provides a continuous coordinate representation through periodic activations (Sitzmann et al., 2020). We represent reliability as a part- and time-conditioned field applied to the complementary paths and evaluate its temporal behavior.

Preference optimization for video diffusion. DPO learns from pairwise preferences without a reward model (Rafailov et al., 2023), and Diffusion-DPO adapts it to diffusion likelihood bounds (Wallace et al., 2024). Video variants align pairs or localize supervision over time and space (Wu et al., 2025; Huang et al., 2026); human-animation methods focus alignment on hands or synthetic detail failures (Wang et al., 2026; Yan et al., 2026). Sign-DPO instead derives the counterfactual and its spatiotemporal support from one phonological perturbation, factorizes supervision by linguistic parameter, and records verifier-equivalent pairs as ties. Its intervention mask also limits the likelihood displacement associated with highly similar pairs (Razin et al., 2025) to the distinguishing region.

3 METHOD

3.1 PROBLEM SETUP AND OVERVIEW

Given reference image I^{ref}, neutral prompt q, and an F-frame driver, we generate $Y = (Y_1, \ldots, Y_F)$ while preserving reference appearance and linguistic motion. Let $\tau_f = (f-1)/(F-1)$ be video time and t diffusion time; Π, π, and a denote projection, model distribution, and phonological factor.

SignDiT retains the reference, text, and face pathways of a pretrained pose-guided video DiT (Cheng et al., 2025) and replaces its pose path with three components (Figure 2): projection-complementary 2D and 3D cues, a part-time reliability field, and phonological-counterfactual Sign-DPO. They determine *what* each path carries, *how much* it is trusted, and *which outputs are semantically correct*.

3.2 PROJECTION-COMPLEMENTARY DUAL-PATH CONDITIONING

Projection fibers. Let $s_f \in \mathcal{S}$ be a fitted articulated 3D body state and let $p_f \in \mathcal{P}$ denote the corresponding 2D keypoints. Conceptually, the camera map $\Pi : \mathcal{S} \to \mathcal{P}$ induces the fiber

$$\mathcal{F}(p_f) = \Pi^{-1}(p_f) = \{s \in \mathcal{S} : \Pi(s) = p_f\}, \tag{1}$$

which contains 3D states that agree after projection and motivates our cue allocation; we do not explicitly represent the complete set. A 2D skeleton specifies image-plane location, scale, and motion, while differences within the fiber include surface depth, front-to-back order, and palm orientation. Conversely, an RGB or shaded rendering of s_f repeats much of the image-plane information already present in p_f. Injecting both therefore asks two independently estimated signals to describe the same degrees of freedom.

We instead construct two *projection-complementary* paths. The first is a skeleton raster $B(p_f)$ obtained from a 2D keypoint estimator. The second is dedicated to cues selected to resolve ambiguities left by projection:

$$\phi_f = \Phi(s_f; \Pi) = \big[N_f,\, D_f^{\mathrm{rel}},\, O_f\big]. \tag{2}$$

Here N_f is a camera-coordinate surface-normal map, $D_f^{\mathrm{rel}}(u) = D_f(u) - D_f(u_{\mathrm{root}})$ is root-relative depth, normalized per sequence by a robust body-scale statistic, and O_f records the front-to-back order of the left hand, right hand, and torso wherever their projections overlap. We explicitly exclude texture, illumination, RGB/CS shading, and a separate silhouette or binary-mask channel. Scalar maps are mapped to fixed three-channel encodings before the frozen VAE; the normal, depth, and order latents are concatenated only inside the fiber path.

Thus the 2D path carries projected motion while the fiber path emphasizes orientation, depth, and occlusion, reducing the duplicated content of a full render; Table 3 compares the alternatives.

Symmetric encoding and additive injection. Let E_{vae} be the frozen video VAE encoder and let $P_{2\mathrm{D}}$ and P_ϕ be trainable patchifiers. The two aligned conditioning-token grids are

$$C^{2\mathrm{D}} = Z_{2\mathrm{D}}(P_{2\mathrm{D}}(E_{\mathrm{vae}}(B(p_{1:F})))), \tag{3}$$

$$C^{\phi} = Z_{\phi}\big(P_{\phi}\big(\mathrm{concat}_{m\in\{N,D,O\}}\, E_{\mathrm{vae}}(m_{1:F})\big)\big), \tag{4}$$

where $Z_{2\mathrm{D}}$ and Z_ϕ are zero-initialized output projections (Zhang et al., 2023). Both grids match the noisy latent tokenization. For frozen noise-patchifier output $H_t = P_z(z_t)$, the first DiT block receives

$$H_t^{\mathrm{in}} = H_t + \widetilde{C}^{2\mathrm{D}} + \widetilde{C}^{\phi}. \tag{5}$$

The tildes denote reliability modulation. Zero initialization makes both paths initially transparent; their shared injection mechanism isolates input content in the C1 ablation.

3.3 Part-conditioned continuous reliability field

Estimator reliability varies across parts and time: the 2D detector loses fingertips under motion blur, while the 3D fit becomes uncertain under hand–face occlusion. After factorization, gating has a precise role as source-specific confidence modulation.

Let $\mathcal{K} = \{\mathrm{body}, \mathrm{left}, \mathrm{right}\}$ be the set of body parts. For each part k and frame f, we compute three observed statistics: mean 2D keypoint confidence $c_{k,f}^{2\mathrm{D}}$, normalized 3D fitting/reprojection residual $r_{k,f}^{\phi}$ (larger is worse), and mean 2D motion magnitude $m_{k,f}$. We linearly interpolate these framewise measurements to functions of video time and write

$$u_k(\tau) = \big[c_k^{2\mathrm{D}}(\tau), r_k^{\phi}(\tau), m_k(\tau)\big]. \tag{6}$$

A sinusoidal representation network (SIREN) (Sitzmann et al., 2020) then defines

$$(h_k^{2\mathrm{D}}(\tau), h_k^{\phi}(\tau)) = g_\eta(e_k, \tau, u_k(\tau)), \tag{7}$$

$$w_k^{2\mathrm{D}}(\tau) = w_{\min} + (1 - w_{\min})\sigma(h_k^{2\mathrm{D}}(\tau)), \tag{8}$$

$$w_k^{\phi}(\tau) = \sigma(h_k^{\phi}(\tau)). \tag{9}$$

where e_k is a learned part embedding. The output biases are initialized so that both weights are close to one. We use $w_{\min} = 0.5$ by default: removing the fiber path leaves a 2D-consistent ambiguity set that the video prior can complete, whereas removing the 2D path discards the primary image-plane trajectory. The asymmetric ranges in Equations 8–9 encode these complementary roles.

Evaluated at the frame times $\tau_{1:F}$, the field yields temporally smoother weights than a framewise MLP (gate TV 0.014 vs. 0.052) and lower hand flicker (Tables 3 and 10).

To apply the scalar weights, we form soft latent-resolution part masks $M_{k,f}$ from dilated convex hulls of the corresponding 2D keypoints. Hand masks take precedence over the body mask and the masks are normalized to a partition of unity. For a token at spatial location i and time τ_f,

$$\widetilde{C}^r_{f,i} = \bar{w}^r_{f,i} C^r_{f,i}, \qquad \bar{w}^r_{f,i} = \sum_{k \in \mathcal{K}} M_{k,f,i} w^r_k(\tau_f), \qquad r \in \{2\mathrm{D}, \phi\}. \tag{10}$$

The same part regions later localize Sign-DPO, coupling reliability-aware injection to semantic credit assignment.

3.4 Stage 1: reconstruction training

We train both patchifiers and zero projections, the reliability field, and self-attention $Q/K/V/O$ LoRA (Hu et al., 2022); the VAE, noise patchifier, remaining DiT blocks, reference path, and face block stay frozen.

Let $z_0 = E_{\mathrm{vae}}(Y)$, let $\epsilon \sim \mathcal{N}(0, I)$, and let $z_t = \alpha(t) z_0 + \sigma(t)\epsilon$ follow the base model's flow schedule (Lipman et al., 2023). The target velocity is $v^*(z_0, \epsilon, t) = \dot{\alpha}(t) z_0 + \dot{\sigma}(t)\epsilon$. With $x = (I^{\mathrm{ref}}, q, p_{1:F}, \phi_{1:F})$, Stage 1 minimizes

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{Y,x,t,\epsilon}\left[\left\| v^*(z_0, \epsilon, t) - v_\theta(z_t, t \mid x) \right\|_2^2\right]. \tag{11}$$

The reliability field g_η is learned end to end through Equation 11. Reference and text conditions are independently dropped with probability 0.05 to retain the base model's classifier-free guidance interface (Ho & Salimans, 2022). In all pose-driven experiments, q is a fixed generic prompt, so semantic content is carried by the driving motion.

3.5 Sign-DPO: phonologically structured preference optimization

Pixel reconstruction gives similar penalties to harmless texture variation and meaning-changing finger or palm errors. We therefore perform a second stage based on direct preference optimization (DPO) (Rafailov et al., 2023), using the Stage 1 model as a frozen reference policy π_{ref}. In contrast to a single video-level preference, Sign-DPO factorizes supervision over

$$\mathcal{A} = \{\mathrm{HS}, \mathrm{LOC}, \mathrm{MOV}, \mathrm{ORI}\}, \tag{12}$$

corresponding to handshape, location, movement, and palm/finger orientation. These four manual factors define the lexical preference space optimized in Stage 2.

Conditional counterfactual pairs. For a training condition x with target gloss g, we select one factor $a \in \mathcal{A}$ and apply a factor-specific edit

$$x^{(a)} = \mathrm{Perturb}_a(x), \qquad Y^+ \sim \pi_{\mathrm{ref}}(\cdot \mid x), \quad Y^- \sim \pi_{\mathrm{ref}}(\cdot \mid x^{(a)}). \tag{13}$$

We use matched reference images, generic prompts, lengths, samplers, and random seeds, so the candidate videos differ as little as possible outside the edited factor. Location and movement edits modify the 2D path; orientation edits modify the fiber path; handshape edits can modify both (Appendix A). Accepted edits correspond to attested lexical minimal pairs or fluent-signer-validated contrasts, keep non-target phonological annotations within pre-specified tolerances, and satisfy kinematic and visual-quality thresholds. The resulting negative expresses a different lexical item at matched visual quality.

Although Y^- is generated from $x^{(a)}$, both videos are scored during DPO under the *same target condition* x: $\pi_\theta(Y^+ \mid x)$ versus $\pi_\theta(Y^- \mid x)$. The counterfactual condition provides a controlled mechanism for producing a fluent, visually plausible semantic negative scored for fidelity to x.

Let $V(g \mid Y)$ be a frozen sign-recognition verifier and let $g^{(a)}$ be the gloss induced by the attested counterfactual. We retain a pair only if

$$\arg\max_{g'} V(g' \mid Y^+) = g, \qquad \arg\max_{g'} V(g' \mid Y^-) = g^{(a)}, \tag{14}$$

and both predictions exceed a confidence threshold. If the verifier assigns the same gloss to both candidates, if the target log-probabilities differ by less than $\varepsilon_{\mathrm{tie}}$, or if either candidate is uncertain, we

record a tie and exclude it from strict-pair training. The retained semantic margin is

$$\delta_a = \gamma \left[\log V(g \mid Y^+) - \log V(g \mid Y^-) - \varepsilon_{\text{tie}}\right]_+, \tag{15}$$

with gradients stopped through V; an independent recognizer evaluates final videos.

Phonology-localized credit. The edit in Equation 13 also defines where the two videos are allowed to affect the preference objective. We construct a latent spatiotemporal mask $M_a = S_a \otimes T_a$. The temporal component T_a contains frames in which x and $x^{(a)}$ differ, with a small temporal dilation. The spatial component reuses the part regions from Equation 10: hands for HS and ORI, hand plus contact locus for LOC, and the swept hand trajectory for MOV. Thus both the negative and its credit-assignment mask follow from the same intervention. For a tensor A, define the normalized masked norm

$$\|A\|^2_{M_a} = \frac{\sum_j M_{a,j} A_j^2}{\sum_j M_{a,j} + \varepsilon}. \tag{16}$$

Shared background, appearance, and unchanged motion receive no direct preference gradient.

Factorized flow-DPO objective. For clarity, the exact sequence-level DPO advantage for a retained pair is

$$A_\theta(Y^+, Y^-; x) = \log \frac{\pi_\theta(Y^+ \mid x)}{\pi_{\text{ref}}(Y^+ \mid x)} - \log \frac{\pi_\theta(Y^- \mid x)}{\pi_{\text{ref}}(Y^- \mid x)}. \tag{17}$$

Following diffusion preference optimization (Wallace et al., 2024), we use a matched-noise denoising surrogate. For the same diffusion time t and Gaussian noise ϵ in both elements of a pair, define

$$d_\theta(Y; x, t, \epsilon, M_a) = \|v^* - v_\theta(z_t, t \mid x)\|^2_{M_a}, \tag{18}$$

$$\Delta^a_\theta(Y; x, t, \epsilon) = d_\theta(Y; x, t, \epsilon, M_a) - d_{\text{ref}}(Y; x, t, \epsilon, M_a), \tag{19}$$

$$\widehat{A}^a_\theta = \omega(t) \left[\Delta^a_\theta(Y^-; x, t, \epsilon) - \Delta^a_\theta(Y^+; x, t, \epsilon)\right], \tag{20}$$

where $\omega(t) > 0$ is the schedule-dependent weight. The sign in Equation 20 is chosen so that reducing the preferred sample's error relative to the reference increases the advantage. The final objective is

$$\mathcal{L}_{\text{Sign-DPO}} = \sum_{a \in \mathcal{A}} \lambda_a \, \mathbb{E}_{\mathcal{D}_a, t, \epsilon} \left[-\log \sigma\Big(\beta \widehat{A}^a_\theta - \delta_a\Big)\right], \tag{21}$$

where $\mathcal{D}_a$ contains only valid, non-tied pairs for factor a. We use equal λ_a by default, preventing factors with more easily mined pairs from dominating. The positive margin δ_a requires a larger policy advantage for verifier-clear semantic contrasts. Table 3 reports the compact comparison, while Table 11 tests negative construction, localization, factorization, the semantic margin, and tie handling.

3.6 Stage 2 and inference

Stage 2 initializes π_θ and frozen π_{ref} from the same Stage 1 checkpoint. Equation 21 updates the self-attention LoRA and condition adapters while keeping the reliability field, pretrained backbone, and verifier frozen. The reference-policy terms provide DPO's implicit proximity constraint without a reward model.

At inference, the 2D keypoints, 3D fit, and reliability statistics are computed once from the driver. We evaluate the field at $\tau_{1:F}$, form Equation 5, and run the base sampler with standard classifier-free guidance (Ho & Salimans, 2022). Pair generation, gloss annotations, and the verifier are training-only and add no inference cost.

4 Experiments

We evaluate visual, hand, and semantic fidelity on three sign languages, plus fluent-signer judgments.

4.1 Experimental protocol

Protocols and data. Self-reenactment takes the reference and driver from one test video, with the reference outside the evaluated window, providing aligned ground truth for reconstruction and

Table 1: Fidelity on How2Sign, CSL-Daily, and Phoenix under (a) paired reconstruction and (b) cross-identity evaluation. H-PCK is hand PCK@0.05; Palm is angular error; HH-Occ. is balanced hand–hand occlusion accuracy. Best/second results are bold/underlined; † marks our reimplementation.

(a) Self-Reenactment — Paired Ground Truth

Dataset	Method	SSIM ↑	LPIPS ↓	FVD ↓	H-PCK ↑	Palm ↓	HH-Occ. ↑
How2Sign	SignMimic	0.837	0.134	913	0.384	27.3	0.781
	SignRefine†	0.863	0.106	637	0.492	23.1	0.824
	SCAIL-2	0.872	**0.0942**	468	0.471	8.6	0.886
	Wan2.2-Animate	0.855	0.108	542	0.407	29.4	0.798
	SignDiT (Ours)	**0.890**	0.117	**272**	**0.517**	**5.1**	**0.957**
CSL-Daily	SignMimic	0.861	0.114	763	0.412	23.7	0.803
	SignRefine†	0.879	0.0885	491	0.531	17.2	0.846
	SCAIL-2	0.893	**0.0771**	387	0.508	7.4	0.905
	Wan2.2-Animate	0.872	0.0934	456	0.441	25.8	0.821
	SignDiT (Ours)	**0.919**	0.101	**255**	**0.589**	**2.5**	**0.964**
Phoenix	SignMimic	0.804	0.183	1060	0.312	33.6	0.746
	SignRefine†	0.822	0.154	771	0.397	26.8	0.788
	SCAIL-2	0.834	0.142	638	0.386	18.5	0.863
	Wan2.2-Animate	0.815	0.163	723	0.334	35.9	0.772
	SignDiT (Ours)	**0.892**	**0.130**	**259**	**0.563**	**5.9**	**0.932**

(b) Cross-Identity Reenactment — Distributional Evaluation

Dataset	Method	FID ↓	FVD ↓	ArcFace ↑	H-PCK ↑	Palm ↓	HH-Occ. ↑
How2Sign	SignMimic	75.3	1090	0.384	0.338	31.2	0.752
	SignRefine†	58.4	673	0.523	0.447	24.6	0.794
	SCAIL-2	54.7	**528**	0.561	0.428	**17.3**	0.861
	Wan2.2-Animate	51.6	594	0.582	0.361	35.1	0.776
	SignDiT (Ours)	**36.0**	543	**0.647**	**0.498**	18.6	**0.925**
CSL-Daily	SignMimic	64.8	917	0.428	0.371	27.4	0.768
	SignRefine†	47.6	562	0.547	0.482	20.3	0.812
	SCAIL-2	44.9	483	0.583	0.461	**13.1**	0.874
	Wan2.2-Animate	42.7	519	0.604	0.392	28.7	0.793
	SignDiT (Ours)	**32.0**	**475**	**0.675**	**0.564**	**13.1**	**0.949**
Phoenix	SignMimic	87.6	1210	0.346	0.272	37.8	0.724
	SignRefine†	69.4	847	0.476	0.359	31.3	0.763
	SCAIL-2	62.8	704	0.518	0.357	22.6	0.831
	Wan2.2-Animate	59.5	776	0.551	0.288	39.4	0.748
	SignDiT (Ours)	**35.0**	**513**	**0.629**	**0.523**	**12.7**	**0.913**

ablations. Cross-identity animation pairs a reference from signer A with a driver from signer B; identity metrics use A and pose/semantic metrics use B. Pairing, shape alignment, crop, duration, and seed are shared. We train on the official training splits and evaluate the official test splits of How2Sign (ASL) (Duarte et al., 2021), CSL-Daily (CSL) (Zhou et al., 2021), and Phoenix (DGS) (Camgoz et al., 2018). No test clip enters training; the official splits share signers, while cross-identity references are unseen synthetic identities that do not depict dataset participants (Appendix B).

Baselines. Baselines cover sign-specific transfer (SignMimic (He et al., 2026)), local hand/face adaptation (SignRefine (Pelykh et al., 2026)), 3D control with Bias-Aware DPO (SCAIL-2 (Yan et al., 2026)), and the Wan2.2-Animate backbone (Cheng et al., 2025). Official checkpoints and settings are used when available; † marks our SignRefine reimplementation. Implementation and checkpoint provenance appear in Appendix A.

Metrics and statistics. SSIM/LPIPS (Wang et al., 2004; Zhang et al., 2018), FID/FVD (Heusel et al., 2017; Unterthiner et al., 2019), and ArcFace (Deng et al., 2019) measure reconstruction, realism, and identity; H-PCK plus palm, depth, and occlusion metrics measure articulation. SpaMo on How2Sign (Hwang et al., 2025) and CV-SLT on Phoenix/CSL-Daily (Zhao et al., 2024) provide BLEU-4 back-translation (Papineni et al., 2002) against the full-sentence references. Sentences longer than one 77-frame window are generated in full by segment-wise continuation as in Wan2.2-Animate (Cheng et al., 2025): each window is conditioned on the last frames of the previous one, and 30-fps resampling preserves duration, so no sentence is truncated or time-compressed. Fixed-manifest automatic metrics use one shared per-clip seed; human intervals follow the task-specific procedures in Appendix F.

4.2 Main quantitative results

Table 2: Back-translation BLEU-4 with language-matched models. Best/second generated-video results are bold/underlined; † marks our reimplementation.

Method	Phoenix (DGS) BLEU-4 ↑	How2Sign (ASL) BLEU-4 ↑	CSL-Daily (CSL) BLEU-4 ↑
SignMimic	18.63	6.41	17.92
SignRefine†	22.41	7.86	21.74
SCAIL-2	<u>23.85</u>	7.12	<u>23.38</u>
Wan2.2-Animate	20.72	<u>8.23</u>	20.49
SignDiT (Ours)	**27.64**	**11.45**	**27.13**
Real video	27.43	13.28	28.79

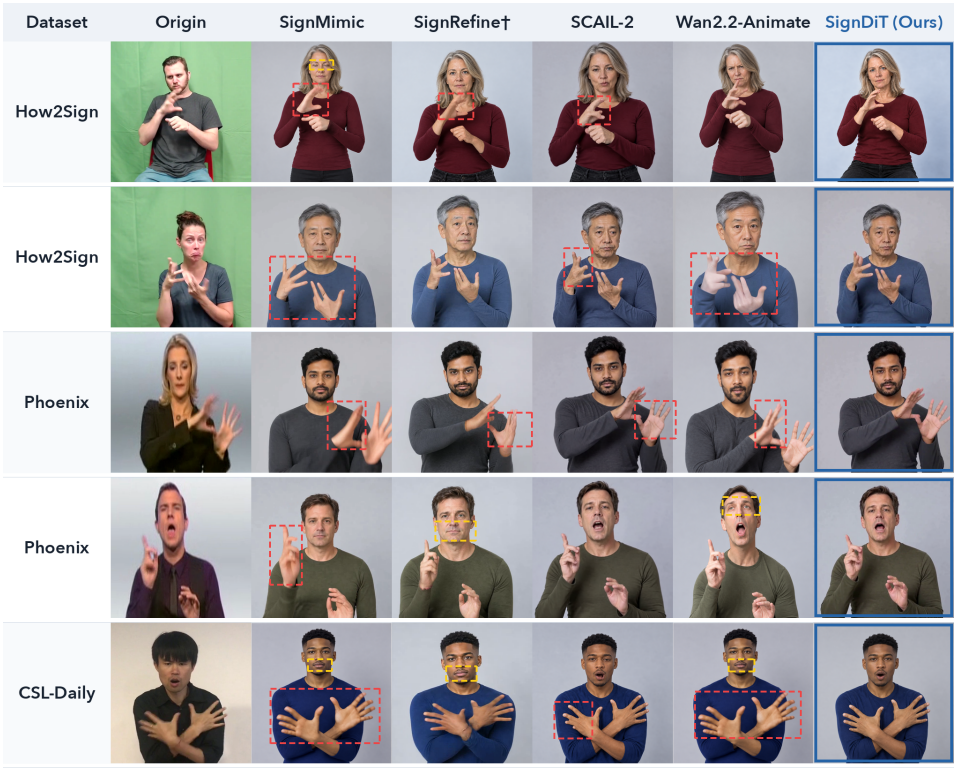

Figure 3: Aligned qualitative comparison across How2Sign, CSL-Daily, and Phoenix; rows share driver timing, and blue borders mark SIGNDIT.

Self-reenactment and cross-identity fidelity. SIGNDIT is best or tied best on 32/36 measures, reducing self-reenactment FVD by 34.1–59.4% and cross-identity FID by 25.1–41.2%. Cross-identity ArcFace, H-PCK, and HH-Occ. improve by 0.065–0.078, 0.051–0.164, and 0.064–0.082, respectively. Exceptions are LPIPS on How2Sign and CSL-Daily, and How2Sign cross-identity FVD/palm error; CSL-Daily palm error is tied.

Back translation. Table 2 measures semantic fidelity by translating each generated video back to spoken-language text. SIGNDIT scores 27.64/11.45/27.13 BLEU-4 on Phoenix/How2Sign/CSL-Daily, beating the strongest generated baselines by 3.79/3.22/3.75 points and remaining within 1.83 points of real video; on blurry, low-resolution Phoenix it slightly exceeds real video (Appendix E).

4.3 QUALITATIVE AND HUMAN EVALUATION

Qualitative comparison. Figure 3 compares representative cases spanning fine handshape, two-hand crossing, open-hand articulation, and hand–face occlusion. All methods share the driver time and signer-centered crop; Appendix H provides aligned temporal strips.

Fluent-signer preference and intelligibility. Eight compensated fluent ASL signers completed institutionally approved, blinded comparisons and recognition on ASL clips. SignDiT was preferred to SignMimic, SignRefine, SCAIL-2, and Wan2.2-Animate in 94/85/78/76% of comparisons (ties:

5/8/14/16%). SignDiT reaches 88% free-response sign accuracy, and participants reported understanding 85% of its sentences (Appendix F). Even when each participant counts as a single sample ($\rho = 1$), the lower bound of the SignDiT-preference interval stays above 50% against SignMimic and SignRefine (Table 13).

4.4 Ablation Studies

Matched How2Sign ablations vary representation (C1), reliability (C2), and preference learning (C3); Appendix D reports all variants; a repeated run differs by 4 FVD, 0.02 H-PCK, and 0.15 BLEU-4.

Component ablations. **C1: Conditioning.** Against a complete render with matched adapters and tokens, fiber maps improve palm error $19° \rightarrow 7°$, FVD $401 \rightarrow 305$, and BLEU-4 $9.04 \rightarrow 9.88$; versus 2D-only, FVD falls from 604 to 305. **C2: Reliability.** The SIREN field gives the best FVD (301) and the lowest flicker (0.012), versus 0.027 for the framewise MLP. **C3: Preference.** Pairs come from WLASL/MS-ASL glosses that differ mainly in one factor; only that factor's local joints are transplanted (Appendix A). Of 15,327 candidates, 7,601 are retained and 2,803 verifier ties are excluded (Table 7). On this pair pool, vanilla Diffusion-DPO raises MP accuracy $0.63 \rightarrow 0.70$ but worsens FVD ($301 \rightarrow 310$), while local masked DPO reaches 0.66. Sign-DPO reaches 0.79 with the best FVD (272) and BLEU-4 (11.45). Although all pairs are ASL, Sign-DPO raises BLEU-4 over Stage 1 by 14.2%, 14.8%, and 5.5% on How2Sign, Phoenix, and CSL-Daily (Table 12).

Table 3: Matched component ablations on How2Sign. MP is minimal-pair discrimination accuracy.

Block	Variant	H-PCK ↑	FVD ↓	Direct metric	BLEU-4 ↑
C1	2D only	0.47	604	Palm: 29	8.27
	2D + full rendering	0.51	401	Palm: 19	9.04
	2D + render + normal/depth	0.52	355	Palm: 12	9.43
	2D + fiber maps	0.52	305	Palm: 7	9.88
C2	No gate	0.50	333	Flicker: 0.021	9.82
	Per-frame MLP	0.49	312	Flicker: 0.027	9.51
	SIREN field	0.50	301	Flicker: 0.012	10.03
C3	Stage 1 only	0.50	301	MP: 0.63	10.03
	Vanilla Diffusion-DPO	0.50	310	MP: 0.70	10.17
	Local random-mask DPO	0.52	298	MP: 0.66	10.52
	Sign-DPO	0.52	272	MP: 0.79	11.45

4.5 Downstream application: word-library-based sign composition

A gloss library retrieves, canonicalizes, and joins one pose clip per gloss, then renders the sequence in one pass; Appendix G gives the protocol.

Table 4: Word-library composition on 50 ASL sequences; naturalness is signer MOS (1–5) on 20, 160 ratings per method.

Method	BLEU-4 ↑	Boundary flicker ↓	ID consistency ↑	Naturalness ↑
Source-video concatenation	8.7	0.038	0.52	1.9
Per-word SignDiT + video concat.	8.1	0.029	0.77	2.6
One-pass Wan2.2-Animate	7.3	0.018	0.75	3.3
One-pass SignDiT w/o SRF	8.5	0.014	0.91	3.9
One-pass SignDiT	9.0	0.010	0.94	4.3

Results. One-pass SignDiT leads with 9.0 BLEU-4, 0.010 flicker, 0.94 identity, and 4.3 naturalness; versus its ungated variant, flicker falls 28.6% while BLEU-4 and mean naturalness rise by 0.5 and 0.4 on the shared manifest.

5 Limitations and Conclusion

Limitations. Our fluent-signer study covers ASL, while CSL and DGS are evaluated with language-matched automatic judges; extending signer studies to these languages is future work. Non-manual markers are carried by the pretrained face pathway, which we keep frozen; jointly modeling them with the manual channels is a natural next step.

Conclusion. SignDiT reframes pose-guided sign generation as structured inference over complementary geometry, learned reliability, and phonological contrast. On three language-specific benchmarks, it links hand fidelity to intelligible, identity-consistent video and lexical composition.

Ethics Statement

Cross-identity sign animation supports accessible content creation, but it can also make a person appear to sign incorrect or falsely attributed content. We use public datasets only under their license terms and animate synthetic reference identities that do not depict dataset participants. Our human-subject study with fluent ASL signers was approved by an Institutional Review Board (IRB) before recruitment, and all participants gave informed consent and were fairly compensated (Appendix F). Generated videos will be labeled as synthetic, impersonation is prohibited in the model card, and high-stakes use requires review by qualified signers (Appendix I).

Reproducibility Statement

The appendix documents data splits, preprocessing, model configuration, training objectives, evaluation metrics, and statistical procedures. Subject to dataset licenses, we will release source code, configuration files, evaluation manifests, and random seeds required to reproduce the experiments, together with the identity-swapped videos generated for all three benchmarks and the synthetic reference identities (Appendix A.4).

AI Use Statement

Large language models were used only to assist with literature search and to polish the writing. GPT Image 2 generated the synthetic reference identities as part of the evaluation protocol (Appendix B). The authors take full responsibility for all content of the paper.

References

Diane Brentari. *A Prosodic Model of Sign Language Phonology*. MIT Press, 1999. ISBN 9780262024457. URL https://mitpress.mit.edu/9780262024457/a-prosodic-model-of-sign-language-phonology/.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Camgoz_Neural_Sign_Language_CVPR_2018_paper.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-Animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. URL https://arxiv.org/abs/2509.14055.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2735–2744,

2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Duarte_How2Sign_A_Large-Scale_Multimodal_Dataset_for_Continuous_American_Sign_Language_CVPR_2021_paper.html.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. SignDiff: Diffusion model for american sign language production. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition*, pp. 1–11, 2025. doi: 10.1109/FG61629.2025.11099440. URL https://doi.org/10.1109/FG61629.2025.11099440.

Changzhou Han, Wanlun Ma, Xi Tang, Kun Hu, Sheng Wen, and Yang Xiang. BoostSLT: Boosting sign language translation via a plug-and-play diffusion-based semantic enhancer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28828–28837, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Han_BoostSLT_Boosting_Sign_Language_Translation_via_a_Plug-and-Play_Diffusion-Based_Semantic_CVPR_2026_paper.html.

Zhewen He, Junyi Yu, Haomian Huang, Zhenhua Li, and Yi Fang. SignMimic: Robust high-quality sign language motion generation via human-shape-oblivious pose transfer guidance. *arXiv preprint arXiv:2609.14122*, 2026. URL https://arxiv.org/abs/2609.14122.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024. doi: 10.1109/CVPR52733.2024.00779. URL https://openaccess.thecvf.com/content/CVPR2024/html/Hu_Animate_Anyone_Consistent_and_Controllable_Image-to-Video_Synthesis_for_Character_Animation_CVPR_2024_paper.html.

Zitong Huang, Kaidong Zhang, Yukang Ding, Chao Gao, Rui Ding, Ying Chen, and Wangmeng Zuo. Mind the generative details: Direct localized detail preference optimization for video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Huang_Mind_the_Generative_Details_Direct_Localized_Detail_Preference_Optimization_for_CVPR_2026_paper.pdf.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3901–3920. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.197. URL https://aclanthology.org/2025.naacl-long.197/.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1459–1469, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Li_Word-level_Deep_Sign_Language_Recognition_from_Video_A_New_Large-scale_WACV_2020_paper.html.

Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. DisPose: Disentangling pose guidance for controllable human image animation. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=AumOa10MKG.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-Sign: Toward unified sign language understanding at scale. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=0Xt7uT04cQ.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Pavlakos_Expressive_Body_Capture_3D_Hands_Face_and_Body_From_a_CVPR_2019_paper.html.

Anton Pelykh, Edward Fish, Ozge Mercanoglu Sincan, and Richard Bowden. SignRefine: Adapting foundational video models for sign language generation. *arXiv preprint arXiv:2609.08496*, 2026. URL https://arxiv.org/abs/2609.08496.

Fan Qi, Yu Duan, Huaiwen Zhang, and Changsheng Xu. SignGen: End-to-end sign language video generation with latent diffusion. In *Computer Vision – ECCV 2024*, volume 15111 of *Lecture Notes in Computer Science*, pp. 252–270, 2024. doi: 10.1007/978-3-031-73668-1_15. URL https://link.springer.com/chapter/10.1007/978-3-031-73668-1_15.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uaMSBJDnRv.

Wendy Sandler and Diane Lillo-Martin. *Sign Language and Linguistic Universals*. Cambridge University Press, 2006. doi: 10.1017/CBO9781139163910. URL https://www.cambridge.org/core/books/sign-language-and-linguistic-universals/3DCD779D499F59838FBC2FC01C0092BC.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5141–5151, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Saunders_Signing_at_Scale_Learning_to_Co-Articulate_Signs_for_Large-Scale_Photo-Realistic_CVPR_2022_paper.html.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908, 2020. doi: 10.1007/s11263-019-01281-2. URL https://doi.org/10.1007/s11263-019-01281-2.

Marshall Thomas, Edward Fish, and Richard Bowden. SignBind-LLM: Multi-stage modality fusion for sign language translation. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science, pp. 262–280, 2026. doi: 10.1007/978-3-032-37550-6_15. URL https://doi.org/10.1007/978-3-032-37550-6_15.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *ICLR 2019 Workshop on Deep Generative Models for Highly Structured Data*, 2019. URL https://openreview.net/forum?id=rylgEULtdN.

Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *British Machine Vision Conference*, pp. 100. BMVA Press, 2019.

Lucas Ventura, Amanda Duarte, and Xavier Giro-i Nieto. Can everybody sign now? exploring sign language video generation from 2d poses. *arXiv preprint arXiv:2012.10941*, 2020. URL https://arxiv.org/abs/2012.10941.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Cong Wang, Zexuan Deng, Zhiwei Jiang, Yafeng Yin, Fei Shen, Zifeng Cheng, Shiping Ge, Shiwei Gan, and Qing Gu. Advanced sign language video generation with compressed and quantized multi-condition tokenization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2658. URL https://papers.nips.cc/paper_files/paper/2025/hash/72295bbd1260b40a0e26952e1531dbfe-Abstract-Conference.html.

Yuanzhi Wang, Xuhua Ren, Jiaxiang Cheng, Bing Ma, Kai Yu, Tianxiang Zheng, Qinglin Lu, and Zhen Cui. Implicit preference alignment for human image animation. *arXiv preprint arXiv:2605.07545*, 2026. URL https://arxiv.org/abs/2605.07545. Accepted at the 43rd International Conference on Machine Learning (ICML 2026).

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling. In *Proceedings of the European Conference on Computer Vision*, 2022.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. DenseDPO: Fine-grained temporal preference optimization for video diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5713. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fa9755043814e7f08d859a286bb83c35-Abstract-Conference.html.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024. URL https://openaccess.thecvf.

com/content/CVPR2024/html/Xu_MagicAnimate_Temporally_Consistent_Human_Image_Animation_using_Diffusion_Model_CVPR_2024_paper.html.

Wenhao Yan, Sheng Ye, Zhuoyi Yang, Jiayan Teng, Zhenhui Dong, Kairui Wen, Xiaotao Gu, Yong-Jin Liu, and Jie Tang. SCAIL: Towards studio-grade character animation via in-context learning of 3d-consistent pose representations. *arXiv preprint arXiv:2512.05905*, 2025. URL https://arxiv.org/abs/2512.05905.

Wenhao Yan, Fengjia Guo, Zhuoyi Yang, and Jie Tang. SCAIL-2: Unifying controlled character animation with end-to-end in-context conditioning. *arXiv preprint arXiv:2606.10804*, 2026. URL https://arxiv.org/abs/2606.10804.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4210–4220, 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/CV4Metaverse/html/Yang_Effective_Whole-Body_Pose_Estimation_with_Two-Stages_Distillation_ICCVW_2023_paper.html.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74896–74910. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25v.html.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19643–19651, 2024. doi: 10.1609/aaai.v38i17.29937. URL https://ojs.aaai.org/index.php/AAAI/article/view/29937.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1325, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhou_Improving_Sign_Language_Translation_With_Monolingual_Data_by_Sign_Back-Translation_CVPR_2021_paper.html.

Jingkai Zhou, Yifan Wu, Shikai Li, Min Wei, Chao Fan, Weihua Chen, Wei Jiang, and Fan Wang. RealisDance-DiT: Simple yet strong baseline towards controllable character animation in the wild. *arXiv preprint arXiv:2504.14977*, 2025. URL https://arxiv.org/abs/2504.14977.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *Computer Vision – ECCV 2024*, volume 15113 of *Lecture Notes in Computer Science*, pp. 145–162, 2024. doi: 10.1007/978-3-031-73001-6_9. URL https://link.springer.com/chapter/10.1007/978-3-031-73001-6_9.

A Implementation and Reproducibility Details

This appendix records the implementation and evaluation protocols.

A.1 MODEL CONFIGURATION

SignDiT starts from the Wan2.2-Animate backbone. The reference-image, text, VAE, and face pathways remain frozen. The trainable components are self-attention LoRA adapters (Hu et al., 2022) on the $Q/K/V/O$ projections, the 2D and fiber patchifiers and their zero-initialized output projections (Zhang et al., 2023), and the SIREN reliability field. The two conditioning streams are encoded separately by the frozen Wan VAE and added to the noisy latent tokens after part-wise reliability weighting. At initialization the zero projections make the new paths transparent, and the SIREN output bias is initialized so that both paths have weight close to one.

Table 5: Implementation settings.

Item	Resolved value
Backbone checkpoint	Wan2.2-Animate-14B
Resolution / frames / frame rate	512×896; 77; 30 fps
VAE spatial / temporal compression	$8\times$ / $4\times$
LoRA rank / α / target modules	64 / 64 / self-attn. $Q/K/V/O$
Fiber-map channel encoding	Normal (n_x, n_y, n_z); replicated depth (d, d, d); fixed three-channel order code. The three RGB-compatible inputs give 9 VAE-input channels and approximately $3 \times 16 = 48$ latent channels.
SIREN layers / width / frequencies	4 affine layers; 256; $\omega_0 = 30$, $\omega_{\text{hidden}} = 3$
2D reliability lower bound $w_{\min}$	0.5
Stage-1 optimizer / LR / schedule	AdamW; 10^{-4}; cosine with 1k-step warmup
Stage-1 steps / batch / grad. accumulation	40k; 1/GPU; 4 (8 GPUs, effective batch 32)
Stage-2 optimizer / LR / steps	AdamW; 10^{-5}; 5k
DPO β / γ / ε_{tie}	0.2 / 0.1 / 0.2
Classifier-free condition dropout	0.05 each for reference and text
Precision / gradient clipping	bfloat16 / 1.0
Sampling solver / steps / CFG scale	DPM++; 25; 1.5
Training hardware / GPU-hours	8 NVIDIA H100 GPUs; 2,000 (Stage 1) + 500 (Stage 2) training GPU-h, excluding pose preprocessing and preference-pair generation

A.2 BASELINE PROVENANCE AND MATCHED INFERENCE

Table 6 records the initialization, original training source, and sign-language adaptation of every comparison system. All methods are evaluated on matched 512×896, 77-frame clips, using their official or paper-specified inference recipes. SignMimic uses its official implementation and checkpoint (He et al., 2026), while SignRefine is our paper-based reimplementation (Pelykh et al., 2026). SCAIL-2 (Yan et al., 2026) and Wan2.2-Animate (Cheng et al., 2025) start from their official checkpoints and receive LoRA fine-tuning on the matched sign-language training set. SignRefine, SCAIL-2, and Wan2.2-Animate are trained for the same number of optimization steps as SignDiT Stage 1 (40k). For every method trained in our pipeline, including SignDiT, the checkpoint is selected on the official validation splits; SignMimic uses its fixed official checkpoint.

Table 6: Baseline provenance and matched sign-language adaptation.

Method	Initialization / weights	Original training source	Sign-language adaptation
SignMimic	Official SignMimic checkpoint	How2Sign, CSL-Daily, and Phoenix official train splits	Official sign-language-trained weights
SignRefine†	Paper-based reimplementation	How2Sign, CSL-Daily, and Phoenix official train splits	Paper-specified local-adapter/LoRA training
SCAIL-2	Official checkpoint `zai-org/SCAIL-2`	Official MotionPair-60K training set	LoRA fine-tuning on the matched sign-language training set
Wan2.2-Animate	Official checkpoint `Wan-AI/Wan2.2-Animate-14B`	Official Wan2.2-Animate training corpus	LoRA fine-tuning on the matched sign-language training set
SignDiT	`Wan-AI/Wan2.2-Animate-14B` + our adapters	How2Sign, CSL-Daily, and Phoenix official train splits	Stage 1 LoRA/adapters + Sign-DPO

Pose preprocessing. DWPose detections and SMPL-X fits (Yang et al., 2023; Pavlakos et al., 2019) are cached once and shared by all compatible methods. Keypoints are mapped through the exact

crop-and-resize transform used for the video. The fiber representation contains surface normals, root-relative depth, and a discrete front-to-back part order; it excludes RGB texture and silhouette channels. Left/right labels are retained through horizontal flips. Failed fits remain in the evaluation set and are flagged rather than silently removed.

Stage 1. Stage 1 uses the official How2Sign, CSL-Daily, and Phoenix training splits with equal dataset sampling probability. We use 77-frame clips sampled at 30 fps. Horizontal flipping is applied with probability 0.5, with left/right labels swapped accordingly; spatial crop/scale and color augmentation are applied with probabilities 0.5 and 0.2, respectively. Only official training splits are used: every validation/test clip is excluded from both training stages. The official splits are clip-disjoint but not signer-disjoint, so self-reenactment test signers also appear in training, whereas cross-identity references are unseen synthetic identities. The objective is the backbone's flow-matching loss; the reliability field receives no direct target. Training runs for 40k steps. Reference-image and text conditions are independently dropped with probability 0.05.

Stage 2. The frozen Stage-1 checkpoint is the reference policy. For $a \in \{\mathrm{HS}, \mathrm{LOC}, \mathrm{MOV}, \mathrm{ORI}\}$, each condition pair differs only in factor a. Candidate pairs are glosses from WLASL and MS-ASL (Li et al., 2020; Vaezi Joze & Koller, 2019) that differ in meaning and whose forms differ mainly in factor a. We extract DWPose and SMPL-X for both signs, align them temporally, normalize skeleton scale, and transplant the local motion of the counterfactual sign into the original condition: HS replaces the finger and palm joints; LOC replaces or translates the wrist and contact position; MOV replaces the hand trajectory; ORI replaces the wrist rotation, palm normals, and corresponding fiber conditions. Non-target body joints, the identity, and frames outside the edited span remain unchanged, and the full condition sequence is re-rendered. The accepted and counterfactual conditions are rendered with identical nuisance variables, and the factor-specific mask is computed before video generation. Pairs that an independent verifier cannot distinguish, or for which both samples receive the same gloss, are recorded as ties and excluded from the strict-pair loss. Stage 2 runs for 5k steps.

Table 7: Sign-DPO preference-pool audit by phonological factor. Counts are measured before factor-balanced Stage-2 sampling.

Factor	Candidates	Accepted	Ties	Uncertain	Other rejects	Gloss contrasts
Handshape (HS)	3,587	1,562	741	629	655	187
Location (LOC)	4,213	2,249	617	533	814	263
Movement (MOV)	5,146	2,873	794	689	790	347
Orientation (ORI)	2,381	917	651	426	387	118
Total	15,327	7,601	2,803	2,277	2,646	915

Both candidates must exceed a calibrated training-verifier confidence of 0.75. An accepted edit changes no non-target coarse factor, changes at most one non-target fine-grained attribute, and has normalized non-target error at most 0.08. We additionally require at least 0.90 valid hand/SMPL-X coverage, normalized keypoint reprojection error at most 0.07, and a joint-limit or temporal-jump violation rate at most 0.06. Visual quality, measured by FAST-VQA (Wu et al., 2022), must exceed the 15th percentile of FAST-VQA scores on the corresponding real training distribution. All acceptance criteria are applied jointly.

Stage 2 first samples the factor uniformly, then a gloss contrast uniformly within that factor, and finally a video realization of the selected contrast. The empirical fraction of samples from each factor is required to remain in $[0.20, 0.30]$. A LOC mask is the union of the hand region and the nearest body-surface contact locus whenever the hand–body distance is at most 0.20 of the torso scale; the contact region is dilated by 0.75 hand widths. A MOV mask rasterizes the continuous wrist and principal-fingertip trajectories with a spatial buffer of 0.75 hand-box scales, or three cells when constructed directly on the latent grid, and is temporally dilated by four video frames on each side. Preference contrasts are drawn from ASL resources and validated by fluent ASL signers.

A.3 Inference and reproducibility

All methods generate matched 512×896, 77-frame clips at 30 fps from the preprocessed driving clip, and metric computation uses the common resize protocol in Appendix C. DWPose and SMPL-X estimates (Yang et al., 2023; Pavlakos et al., 2019) are cached from a single preprocessing manifest

when compatible with a method's interface. SignDiT samples with DPM++ for 25 steps at a CFG scale of 1.5.

Every method is evaluated with one predetermined generation seed per clip. Reference images, text (if used), drivers, frame indices, crop transforms, and generation seeds live in one immutable JSONL manifest. Generated videos are decoded once and stored losslessly for metric computation. All metric programs consume the same file list and emit both per-clip records and aggregates.

A.4 Code and data release

We will open-source the complete SignDiT codebase, covering preprocessing, both training stages, inference, and evaluation, together with configuration files, evaluation manifests, and random seeds. We will also release the identity-swapped data produced in this work: the cross-identity videos generated for How2Sign, CSL-Daily, and Phoenix, together with the 200 synthetic reference identities and the driver–reference assignments of Appendix B. This release is particularly valuable for low-resolution sources such as Phoenix, whose 210×260 broadcast recordings contain motion blur and compression artifacts; the regenerated videos keep the original signing content at 512×896 and 30 fps. Because the synthetic identities do not depict dataset participants, the data can also be shared without exposing signers' faces. We expect this release to benefit the sign-language community as a resource for training and evaluating recognition, translation, and production models. Distribution follows each source dataset's license terms (Appendix I).

B Datasets, Splits, and Evaluation Manifests

The main benchmarks are How2Sign, CSL-Daily, and Phoenix, with results reported separately for ASL, CSL, and DGS. WLASL and MS-ASL provide the isolated-ASL word pool (Li et al., 2020; Vaezi Joze & Koller, 2019). We use only resources that can be legally and reproducibly obtained and do not replace an unavailable resource after inspecting test performance.

Table 8: Datasets and their roles. Metrics are reported separately for each sign language.

Dataset	Domain / language	Role	Training use
How2Sign (Duarte et al., 2021)	Continuous ASL	Primary benchmark; ablations	Official train only
CSL-Daily (Zhou et al., 2021)	Continuous CSL	Main benchmark	Official train only
Phoenix (Camgoz et al., 2018)	Continuous DGS	Main benchmark	Official train only
WLASL (Li et al., 2020)	Isolated ASL	Word library; preference pairs	Conditions only
MS-ASL (Vaezi Joze & Koller, 2019)	Isolated ASL	Word library; preference pairs	Conditions only

Evaluation-manifest scale. We evaluate the complete official test splits: 2,322 How2Sign clips, 1,176 CSL-Daily clips, and 642 Phoenix clips. Both self-reenactment and cross-identity evaluation therefore contain 2,322, 1,176, and 642 units, respectively. Each unit contains 77 frames. Per protocol, this gives 178,794/90,552/49,434 generated frames and, using every overlapping 16-frame window, 143,964/72,912/39,804 FVD windows for How2Sign, CSL-Daily, and Phoenix. Each generated unit is matched one-to-one with a real test clip, so the real and generated sets have equal cardinality.

FID uses the FID-specific InceptionV3 weights from PyTorch-FID ("pt_inception-2015-12-05-6726825d.pth") and its 2048-D final average-pooling activations. Real and generated frames are sampled at identical manifest indices with equal cardinality. FVD uses the Google Research reference implementation with the TF-Hub `deepmind/i3d-kinetics-400/1` I3D model and `RGB/inception_i3d/Mean:0` embeddings. Embeddings from all overlapping 16-frame windows with identical start indices are pooled before one Fréchet computation.

Self-reenactment manifest. For continuous signing, a reference frame is drawn from the same source video but outside the evaluated sentence whenever possible. Each unit is one full test sentence at 30 fps; sentences longer than 77 frames are generated in consecutive 77-frame windows, each conditioned on the last frames of the previous window, following Wan2.2-Animate (Cheng et al., 2025). The manifest stores the original frame IDs so all methods and metrics observe the same content.

Cross-identity manifest. Each test driver is paired once, using a fixed random seed, with a reference drawn from a bank of 200 synthetic identities generated with GPT Image 2. These reference images do not depict dataset participants or named real individuals, and the same driver–reference pairing is used for every method. Every identity is generated from one fixed prompt template, in which {PERSON} describes the identity and {COLOUR} the colour of the top:

> *Photorealistic studio portrait of {PERSON}, facing the camera straight on, medium shot from above the knees up with headroom, arms hanging down so the hands are out of frame. Neutral expression, mouth closed, looking into the lens. Plain fitted {COLOUR} long-sleeved crew-neck top, no logo or pattern, black trousers.*

Keeping the hands out of frame prevents the reference from supplying a hand pose, and the plain clothing and neutral expression keep appearance cues separate from the driven motion. SMPLify-X adapts body shape to the reference while retaining the driver's pose (Pavlakos et al., 2019). The alternative SE(3)-normalized front end is reported as a separate pipeline choice, not selected per example.

Leakage and deduplication. We hash decoded frames, compare near-duplicate visual embeddings, and match source IDs across every training and test manifest. Any removed item, threshold, and manual adjudication is logged.

C Metric Definitions and Statistical Analysis

Let $\hat{v}$ denote a generated video, v a pixel-aligned target when it exists, r the reference image, and d the driving pose sequence. SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) are averaged over aligned valid frames only. FID (Heusel et al., 2017) uses 2048-D FID-InceptionV3 final-pool features from the same number of uniformly sampled frames for every method. FVD (Unterthiner et al., 2019) uses official Kinetics-400 I3D embeddings from overlapping 16-frame windows with identical start indices. All videos are resized to the common evaluation size before feature extraction.

For keypoint type j with evaluator prediction $\hat{q}_{tj}$ and driver target q_{tj}, hand PCK (H-PCK) at threshold τ is

$$\mathrm{PCK}_\tau = \frac{\sum_{t,j} a_{tj}\, \mathbf{1}[\|\hat{q}_{tj} - q_{tj}\|_2 < \tau s_t]}{\sum_{t,j} a_{tj}},$$

where s_t is the longer side of the driver hand box and a_{tj} requires a valid target and evaluator detection. Coverage is $\sum a_{tj} / \sum \mathbf{1}[q_{tj}\ \text{valid}]$; Table 10 reports it for the reliability ablations.

For fitted palm normals $\hat{n}_t$ and driver normals n_t, angular error is

$$e_{\mathrm{palm}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \frac{180}{\pi} \arccos\big(\mathrm{clip}(\hat{n}_t^\top n_t, -1, 1)\big)\,.$$

We report left and right hands separately before their macro-average. Line-of-sight articulation error compares root-relative fitted MANO joint coordinates projected onto the camera viewing axis, after root and scale alignment. Hand–hand occlusion-order accuracy (HH-Occ.) is evaluated only on frames in which the projected fitted left- and right-hand meshes overlap in the driver. Let $y_t \in \{L, R\}$ denote the front hand in the driver and $\hat{y}_t \in \{L, R\}$ the front hand obtained by refitting the generated frame. With $\mathcal{V}_L = \{t : y_t = L\}$ and $\mathcal{V}_R = \{t : y_t = R\}$, we compute the balanced accuracy

$$\text{HH-Occ.} = \frac{1}{2}\left(\frac{\sum_{t \in \mathcal{V}_L} \mathbf{1}[\hat{y}_t = L]}{|\mathcal{V}_L|} + \frac{\sum_{t \in \mathcal{V}_R} \mathbf{1}[\hat{y}_t = R]}{|\mathcal{V}_R|}\right).$$

Thus, always predicting the same ordering scores 0.5, whereas reproducing every valid driver ordering scores 1.0. Frames with invalid fits or an ambiguous driver depth ordering are excluded. The minimum projected-overlap area and depth-tie tolerance are fixed on the training split.

Hand-region temporal flicker is the variance of motion-compensated residuals between adjacent frames inside the union of driver hand masks. Boundary flicker uses the same definition in a ± 4-frame window around a stitch. Identity consistency is the mean adjacent-segment cosine similarity of full-body DINO features (Caron et al., 2021); ArcFace (Deng et al., 2019) is reported only when a face detector succeeds.

Semantic metrics are computed by frozen judges that never train SignDiT. Text BLEU-4 uses the official reference translations and tokenization (Papineni et al., 2002). Minimal-pair discrimination counts a pair as correct only when the judge assigns each output to its corresponding gloss.

Primary comparisons and metrics are declared before reading results. Automatic metrics are reported as point estimates on the fixed manifest and shared per-clip generation seeds. Human intervals follow the endpoint-specific procedures reported in Appendix F.

D Complete Component Ablations

All component ablations use the same How2Sign self-reenactment manifest, training seed, and per-clip generation seeds. C1 is an independently trained matched sweep that holds the default SIREN architecture fixed while varying the conditioning representation. C2 fixes the selected fiber representation and independently compares reliability mechanisms. C3 starts from the selected C2 SIREN checkpoint, so its Stage-1 row matches the C2 SIREN row exactly. Stage-1 ablations are compared before preference optimization. We report both the component's direct metric and guardrail metrics. The selected configuration therefore appears as two separate training runs, the C1 fiber-map row and the C2 SIREN row; their difference (4 FVD, 0.02 H-PCK, 0.15 BLEU-4) gives the run-to-run variation, which the FVD and BLEU-4 effects discussed in Section 4.4 exceed. The 2D-only, full-render, and additive render/normal/depth variants adopt the conditioning inputs of Wan2.2-Animate (Cheng et al., 2025), RealisDance-DiT (Zhou et al., 2025), and Champ (Zhu et al., 2024), respectively. The keypoint-confidence hard mask is modeled on MimicMotion (Zhang et al., 2025). The preference variants are modeled on Diffusion-DPO (Wallace et al., 2024), LocalDPO (Huang et al., 2026), SCAIL-2 (Yan et al., 2026), and IPA (Wang et al., 2026), respectively. Each variant replaces only the named component inside the SignDiT pipeline, keeping its adapters, training budget, and remaining components, so the variants isolate that component rather than reproduce the complete systems in Table 1.

Table 9: C1: representation supplied to the second conditioning path.

Variant	FVD ↓	H-PCK ↑	Palm ↓	Depth ↓	HH-Occ. ↑	BLEU-4 ↑
2D only	604	0.47	29	0.28	0.64	8.27
Fiber only	620	0.39	11	0.13	0.78	8.03
2D + full SMPL-X rendering	401	0.51	19	0.20	0.69	9.04
2D + render + normal + depth	355	0.52	12	0.14	0.77	9.43
2D + fiber maps	305	0.52	7	0.09	0.91	9.88
− occlusion-order map	340	0.52	9	0.10	0.64	8.62
− normal map	380	0.48	19	0.10	0.85	9.21
− relative-depth map	355	0.52	10	0.20	0.80	8.47

Table 10: C2: reliability injection with C1 fixed. Gate TV is the temporal total variation of the predicted part weights; Coverage is valid DWPose hand-keypoint coverage.

Variant	FVD ↓	H-PCK ↑	Flicker ↓	Gate TV ↓	BLEU-4 ↑	Coverage
No reliability gate	333	0.50	0.021	—	9.82	0.93
Keypoint-confidence hard mask	352	0.50	0.033	0.090	9.36	0.91
Independent per-frame MLP	312	0.49	0.027	0.052	9.51	0.94
SIREN reliability field	301	0.50	0.012	0.014	10.03	0.96
SIREN, symmetric gates	312	0.51	0.016	0.017	9.84	0.95

E Semantic-Judge Validation

The verifier used to construct or filter preference pairs is disjoint from all evaluation judges. The training verifier is Uni-Sign (Li et al., 2025b), using its official release. Evaluation judges are selected from official releases that provide source code, a downloadable trained checkpoint, and an executable evaluation path as of September 19, 2026. Under this reproducibility constraint, we use SpaMo for How2Sign (Hwang et al., 2025) and the official CV-SLT checkpoints for Phoenix and CSL-Daily (Zhao et al., 2024). We run each checkpoint on generated and real videos from the same fixed manifest using identical preprocessing, decoding, tokenization, and BLEU computation. The resulting real-video BLEU-4 scores are 13.28 on How2Sign, 27.43 on Phoenix, and 28.79 on CSL-Daily. Phoenix source videos are low-resolution (210×260) broadcast recordings with visible motion blur and

Table 11: C3: preference optimization with C1+C2 fixed. Disp. is normalized chosen-sample likelihood displacement relative to the Stage-1 reference, with higher values indicating stronger preference. "Retain ties" trains tied pairs as ordinary pairs, preferring the original-gloss video with zero semantic margin.

Variant	FVD ↓	H-PCK ↑	BLEU-4 ↑	MP acc. ↑	Disp. ↑
Stage 1 only	301	0.50	10.03	0.63	—
Vanilla Diffusion-DPO	310	0.50	10.17	0.70	−0.14
Local random-mask DPO	298	0.52	10.52	0.66	−0.01
Degraded-hand negatives	295	0.50	10.58	0.67	−0.02
Hand-region implicit alignment	290	0.47	10.64	0.71	0.01
Sign-DPO − factorization	285	0.52	10.71	0.74	0.08
Sign-DPO − semantic margin	280	0.49	10.86	0.76	0.09
Sign-DPO, retain ties	290	0.52	10.69	0.73	0.04
Sign-DPO	272	0.52	11.45	0.79	0.14

compression artifacts during fast hand movements. Generated videos largely avoid these degradations (Figures 3 and 4), consistent with the slightly higher Phoenix score (+0.21 BLEU-4). Paper-only systems and repositories without released inference weights are not used, irrespective of reported leaderboard score. A judge is not used outside the language and label set on which it was calibrated.

Table 12: Per-language effect of Sign-DPO using the same Stage-1 checkpoint, manifest, and language-matched judge. This isolates Stage 2 from the complete-system comparison in Table 2.

Dataset	Stage-1 BLEU-4	Sign-DPO BLEU-4	Δ
How2Sign	10.03	11.45	+1.42
CSL-Daily	25.71	27.13	+1.42
Phoenix	24.08	27.64	+3.56

All Sign-DPO preference pairs are constructed from ASL resources, yet Table 12 shows gains in all three languages. The four optimized manual parameters are shared across sign languages (Brentari, 1999; Sandler & Lillo-Martin, 2006), which is consistent with the hand-localized preference signal transferring beyond the ASL lexicon. Relative to Stage 1, BLEU-4 rises by 14.2% on How2Sign, 14.8% on Phoenix, and 5.5% on CSL-Daily.

F IRB-Approved Human-Subject Study

The study received Institutional Review Board approval before recruitment. We recruited and compensated eight fluent ASL signers as evaluators, and all provided informed consent. Institution-specific protocol identifiers and approval details are withheld during double-blind review.

Tasks. Each participant completed four tasks; all stimuli were ASL and were sampled at random rather than hand-picked. *Isolated-sign recognition* showed 40 SignDiT videos of isolated WLASL signs (Li et al., 2020); participants reported the gloss as a free response, without candidate options; a response counted as correct only if it matched the target gloss. *Sentence comprehension* showed 20 generated SignDiT sentences and asked participants whether they understood the meaning. *Naturalness* used 20 of the 50 composed sequences in Table 4; each sequence was rendered by all five composition methods, and participants rated each rendering from 1 to 5 by whether one person appeared to sign the full sentence naturally. *Pairwise preference* presented 30 matched SignDiT–baseline pairs per baseline and recorded SignDiT / Tie / baseline preference. Participants were instructed to judge sign correctness first and to use visual quality only when both videos were equally correct. Pairs were drawn from the How2Sign test manifest; no pair was shown to more than one participant, and no participant saw the same source clip in comparisons against different baselines. All tasks were blinded: method names were never shown, trial order was randomized, and left/right placement in pairwise trials was randomized.

Analysis. Trials are nested within eight participants. Exact video pairs were not repeated across participants, which removes clustering by identical pair; to account for within-participant dependence, we assess the preference results under a design effect $1 + (m - 1)\rho$, where $m = 30$ trials per participant and ρ is the within-participant correlation; the effective sample size is $240/(1 + 29\rho)$, and break-even values use the reported percentages under an exchangeable within-participant correlation

Table 13: Pairwise preference from eight fluent ASL signers. Each trial records preference for SignDiT, a tie, or preference for the baseline. Trial-level CI is the Wilson score interval for the SignDiT-preference rate, which treats trials as independent. Break-even ρ is the within-participant correlation at which the design-effect-adjusted lower bound reaches 50%; "none" means it stays above 50% for every $\rho \in [0, 1]$.

Baseline	Participants	Trials / participant	Included trials	Prefer SignDiT	Tie	Prefer baseline	Trial-level 95% CI	Break-even ρ
SignMimic	8	30	240	94%	5%	1%	[90, 96]%	none
SignRefine	8	30	240	85%	8%	7%	[80, 89]%	none
SCAIL-2	8	30	240	78%	14%	8%	[72, 83]%	0.64
Wan2.2-Animate	8	30	240	76%	16%	8%	[70, 81]%	0.55

Table 14: SignDiT human intelligibility and naturalness outcomes. Isolated signs use free responses; intervals are trial-level Wilson intervals; naturalness is the SignDiT mean rating.

Endpoint	Participants	Trials / participant	Included trials	Estimate	Trial-level 95% CI
Isolated-sign accuracy	8	40	320	88%	[84, 91]%
Sentence comprehension	8	20	160	85%	[79, 90]%
Stitched-sentence naturalness	8	20	160	4.3 / 5	–

model. Against SignMimic and SignRefine, the adjusted lower bound remains above 50% even at $\rho = 1$, where the effective sample size equals the number of participants. Against SCAIL-2 and Wan2.2-Animate, it falls to 50% only when ρ exceeds 0.64 and 0.55, respectively. At $\rho = 0.2$, the SignDiT preference intervals are [81, 98]%, [70, 93]%, [62, 89]%, and [60, 87]%. No sessions or trials were excluded: all eight participants contributed every trial counted in Tables 13 and 14.

G Word-Library-Based Sign Composition

We construct an ASL motion library $\mathcal{L} = \{(g_i, P_i)\}_{i=1}^{N}$ from gloss-labeled isolated-sign videos in WLASL and MS-ASL (Li et al., 2020; Vaezi Joze & Koller, 2019), where g_i is a lexical gloss and P_i its extracted pose sequence. The curated vocabulary retains entries with stable pose extraction, clear lexical labels, and signer-validated variants. We then create 50 three-to-six-gloss sequences from this vocabulary and validate their lexical content with fluent ASL signers. The released manifest records glosses, candidate clips, selected clips, signers, and normalization parameters.

For each requested gloss, deterministic retrieval selects a pose clip from $\mathcal{L}$. DWPose and SMPL-X (Yang et al., 2023; Pavlakos et al., 2019) supply projected and fiber conditions; root motion, body scale, and coordinate frames are canonicalized to the target avatar. Adjacent lexical motions are joined with fixed-duration pose-space transitions, and the complete sequence is rendered in one pass from a single reference image.

H Additional Qualitative Results

Figure 4 shows nine sequences: three each from How2Sign, CSL-Daily, and Phoenix. Each row contains a reference identity, five original frames, and the corresponding time-aligned SignDiT outputs.

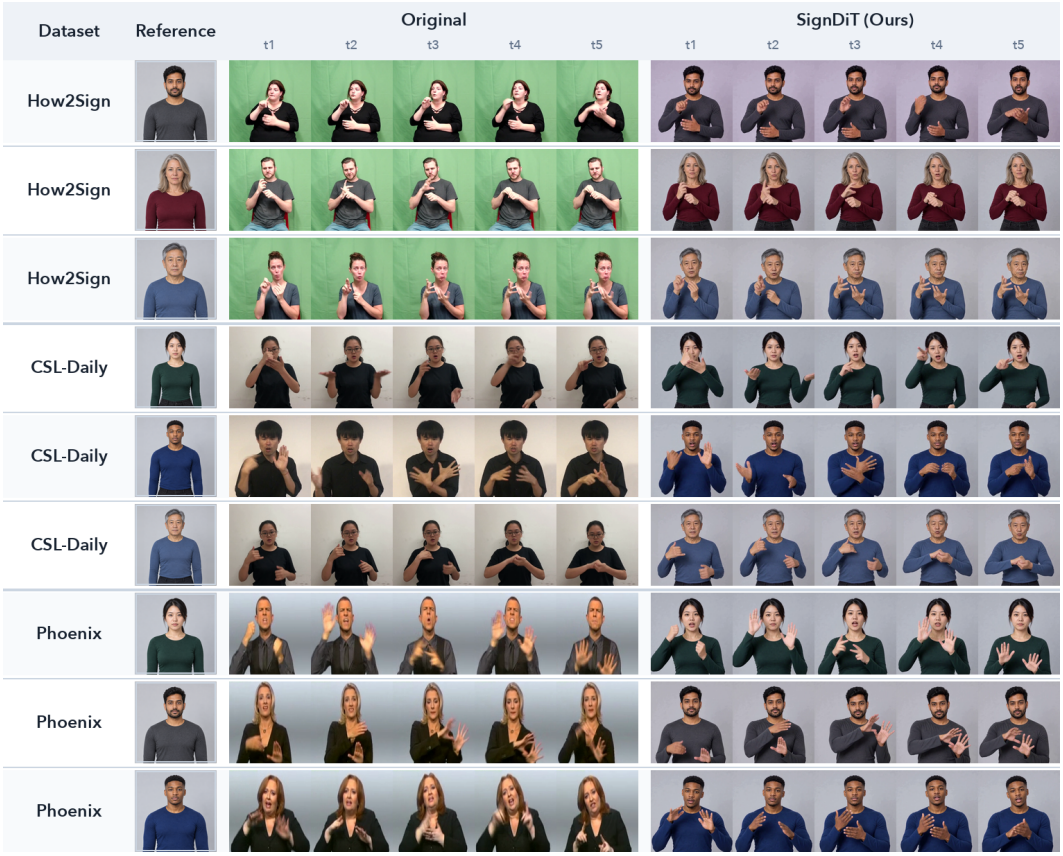

Figure 4: Temporal qualitative results on How2Sign, CSL-Daily, and Phoenix. Each row shows one reference image, five original frames, and the corresponding time-aligned SignDiT outputs.

I Ethics, Data Governance, and Intended Use

Sign-language video is not merely generic human-motion footage: it can contain faces, biometric cues, disability or language-community information, and linguistic content. We use each dataset only under its license and consent terms, document whether derivative pose data may be stored or shared, and do not redistribute source video through our manifests. Public release will contain source identifiers or hashes only where permitted. Requests for removal will be propagated to cached poses, preference pairs, and generated examples.

Official training and test splits remain clip-disjoint, and no test clip enters either training stage. Cross-identity references use synthetic identities that do not depict dataset participants or named real individuals. Access to identifiable raw videos and face embeddings is restricted to approved researchers; derived biometric embeddings are used only transiently for evaluation and are deleted according to the data-management schedule. Logs and qualitative figures are checked for names, faces not covered by the source consent, filesystem paths, and other identifying metadata. Demographic performance is reported only for consented, sufficiently large groups and never from automatic sensitive-attribute inference.

The system may support accessible content creation, avatar-based communication, and data augmentation, but it may also generate a fluent-looking person signing incorrect, offensive, or falsely attributed content. Visual realism must not be taken as linguistic validity. We will release a model card documenting supported sign languages, known failure modes (especially non-manual markers), intended research use, and prohibited impersonation or deceptive deployment. Generated demonstrations will be labeled as synthetic and will not mimic an identifiable person without authorization. High-stakes communication, translation, education, or medical/legal use requires review by qualified Deaf or fluent signers.

The completed human study followed the safeguards in Appendix F; all reported results were collected after IRB approval and informed consent. Community expertise is treated as substantive evaluation expertise rather than as a post-hoc preference check. Institution-specific approval details are withheld during double-blind review.