

SignNet-1M: Large-Scale Multilingual Sign Language Video Dataset with Downstream Benchmarks

Zhewen He¹, Junyi Hu¹, Haomian Huang¹, Zhenhua Li^{1,2}, Yu-Shen Liu³, and Yi Fang^{1,2*}

¹ New York University Abu Dhabi, United Arab Emirates

² ChatSign Technology, United Arab Emirates

³ Tsinghua University, China

zh3510@nyu.edu, yf23@nyu.edu

Abstract. Sign language models are typically trained on datasets captured under constrained conditions, with limited viewpoint, background, and signer-identity diversity, leading to poor robustness under real-world distribution shifts. We introduce **SignNet-1M**, a large-scale augmented dataset spanning ASL, CSL, and German Sign Language (DGS). SignNet-1M synthesizes realistic variations along three axes: (i) *novel-view* rendering (rotation and zoom) via 3D Gaussian Splatting (3DGS), (ii) *scene/identity* editing via diffusion models for background replacement and signer substitution while preserving sign motion and linguistic content, and (iii) *post-rendering* augmentations that emulate capture and compression artifacts (e.g., geometric transforms, photometric shifts, mild temporal resampling, and compression) to better match in-the-wild recordings. Beyond data release, we provide a unified benchmark suite across downstream tasks (e.g., translation and recognition) and ablations that isolate each augmentation component. Experiments across backbones show that training with SignNet-1M consistently improves generalization under cross-view, cross-background, cross-identity, and post-rendering shifts, while maintaining strong in-distribution performance. The dataset, released augmentation components, metadata, and benchmark resources are available at <https://signnet.chatsign.ai/>.

Keywords: sign language · dataset · augmentation · 3DGS · diffusion

1 Introduction

Sign language translation (SLT) has advanced rapidly on standard benchmarks such as Phoenix14T [1], How2Sign [5] and CSL-Daily [34], yet these datasets are captured under constrained conditions—near-frontal cameras, studio backgrounds, and few signers—that poorly represent real-world deployment. State-of-the-art models such as SpaMo [9] and UniSign [13] are consequently

* Corresponding author.

brittle to *viewpoint*, *background*, signer-appearance shifts, and common image-/time-domain corruptions (e.g., photometric and compression artifacts). For example, SpaMo trained on Phoenix14T shows a substantial BLEU-4 drop under viewpoint augmentation, highlighting a critical coverage blind spot of standard benchmarks (Sec. 6.2).

Conventional pixel-level augmentations (cropping, flipping, color jitter) cannot synthesize the structured, 3D-consistent variations that characterize viewpoint/background/identity shifts. Recent advances in 3D Gaussian Splatting (3DGS) [10] and diffusion-based video editing [6] now make it feasible to produce geometry-aware viewpoint, scene, and appearance variations at scale while preserving temporal consistency—opening the door to *structure-aware* sign language data augmentation.

We introduce **SignNet-1M**, a large-scale augmented multilingual sign language dataset of ~ 1 M clips across three languages (ASL, DGS, CSL) drawn from seven source corpora. Each clip is augmented along three axes: (i) *viewpoint* via 3DGS novel-view rendering (GUAVA [31]), (ii) *background* via diffusion-based scene editing (FlowPortal [6] + IC-Light [33]), and (iii) *identity* via cross-reenactment signer substitution. We additionally apply a lightweight *post-rendering augmentation* stage (standard image-space transforms plus mild temporal resampling) to further broaden clip diversity. We pair the dataset with a four-setting evaluation protocol that quantifies both the coverage blind spot of existing benchmarks and the training value of augmented data. Our contributions are:

- **SignNet-1M dataset**: ~ 1 M augmented clips spanning ASL, DGS, and CSL with controllable viewpoint, background, and identity variations from seven source corpora.
- **Scalable augmentation pipeline**: 3DGS novel-view rendering + diffusion-based scene/identity editing, plus lightweight post-rendering augmentations, engineered to preserve sign dynamics and linguistic labels.
- **Benchmark protocol**: a unified evaluation framework with severity-stratified robustness analysis and ablations that reveal where SLT models fail and how augmented training recovers performance.

2 Related Work

2.1 Sign language datasets and SLT pipelines.

SLT benchmarks have expanded in scale and language coverage.

Phoenix14T [1] established the standard DGS setting with aligned gloss and translation supervision. ASL benchmarks such as How2Sign [5], OpenASL [20], and YouTube-ASL [23], together with CSL-Daily [34] and CSL-News [13], expanded the language and domain coverage. Signer-independent collections (e.g., FluentSigners-50 [16]) further highlight growing interest in generalization across identities. However, most training data remain biased toward constrained capture—near-frontal views, limited backgrounds, and restricted

signer pools—making robustness to real-world shifts hard to assess in a unified manner.

SLT modeling has co-evolved with these datasets. Early systems largely follow a gloss-supervised paradigm, translating videos via an intermediate gloss sequence [1]. Transformer-based SLT strengthened sequence modeling, e.g., Sign Language Transformers [2] and STMC-Transformer [28]. More recently, methods exploit stronger vision encoders and vision-language alignment: SpaMo performs gloss-free translation with prompt-based LLM decoding [9], while Uni-Sign incorporates pose/skeleton streams as explicit motion priors and a unified pre-training/fine-tuning recipe [13]. Yet most gains are reported on i.i.d. splits drawn from the same recording setups, leaving factor-specific robustness (viewpoint, background, signer identity) under-quantified. SignNet-1M complements this line with controlled, factorized visual shifts and a unified protocol for diagnosing robustness.

2.2 Data augmentation and robustness under distribution shifts.

Data augmentation is central to robustness in visual recognition, including policy-based transforms (RandAugment [3]), composition (AugMix [7]), and sample mixing (MixUp [32], CutMix [29]). For video, temporal coherence further constrains valid augmentations, motivating video-specific invariances and actor/scene mixing (e.g., ActorCutMix [35]). However, such pixel/patch-level transforms are insufficient for SLT: sign meaning depends on subtle hand/face motion that can be corrupted by aggressive appearance edits, and real deployments exhibit structured shifts (viewpoint, background, unseen signer) that are difficult to synthesize with frame-wise operations alone.

Sign-language augmentation has mainly targeted data scarcity or alignment rather than factor-controlled shifts. Sign Back-Translation (SignBT) [34] leverages monolingual text to enrich SLT pairs, and other cross-modality strategies introduce auxiliary cues, but they do not provide explicit, controllable variation along viewpoint/background/identity axes at scale. In contrast, SignNet-1M generates factorized, controllable visual variations while preserving motion and linguistic annotations, enabling severity-stratified robustness re-training and evaluation.

2.3 3D-aware rendering and diffusion-based video editing.

Controllable generation enables structured variations beyond conventional augmentation. 3D Gaussian Splatting (3DGS) [10] supports efficient novel-view synthesis with explicit camera control. GUAVA [31] builds animatable upper-body avatars via SMPL-X [18] and FLAME [12], enabling geometry-consistent viewpoint manipulation and cross-identity reenactment through disentangled identity and motion parameters.

Diffusion-based video generation/editing has also progressed rapidly, including Tune-A-Video [25], Text2Video-Zero [11], and Video-P2P [15]. For

practical scene editing, FlowPortal [6] performs temporally consistent relighting/background replacement via residual-corrected flow in latent space, while IC-Light [33] provides illumination-aware conditioning. SignNet-1M assembles these components into a sign-language-tailored augmentation pipeline: 3D-aware avatars provide controllable viewpoint/identity variation, and diffusion editing introduces diverse, temporally consistent background changes—together yielding factorized shifts that can be systematically benchmarked.

3 Methodology

We construct SignNet-1M by applying controlled factor shifts to each source clip along three axes—*viewpoint*, *background*, and *signer identity*—followed by lightweight *post-rendering* augmentations. For notational convenience, we denote a clip as $\mathcal{V} = \{v_t\}_{t=1}^T$ when needed. Fig. 1 summarizes the procedure, and Fig. 2 illustrates typical outcomes (a–h).

Concretely, we optionally replace the background to create an alternative stream $\mathcal{V}^{\text{bg}}$ (Stage 1), track the signer to a parametric representation and render novel views and cross-identity reenactment on both $\mathcal{V}$ and $\mathcal{V}^{\text{bg}}$ (Stage 2), and finally apply lightweight post-rendering augmentations that preserve linguistic annotations (Stage 3), together spanning all three factor axes (Fig. 2a–h).

3.1 Novel-View Rendering via 3DGS

Parametric tracking. To enable 3D-aware view synthesis while preserving the original signing motion, we first track each clip to a compact parametric representation. EHM-Tracker [30] fits SMPL-X [18] + FLAME [12] per frame and factorizes clip-level identity from time-varying motion:

$$\Theta_t = \underbrace{(\beta_b, \beta_f, \Delta J)}_{\Theta^{\text{id}}}, \underbrace{(\theta_t^b, \theta_t^h, \theta_t^{\text{jaw}}, \psi_t)}_{\Theta_t^{\text{mot}}}, \quad (1)$$

where the identity block Θ^{id} gathers the SMPL-X body shape β_b , FLAME facial shape β_f , and personalized joint offsets ΔJ (constant within a clip), and the motion block Θ_t^{mot} gathers the per-frame body, hand, and jaw poses $\theta_t^b, \theta_t^h, \theta_t^{\text{jaw}}$ and facial expression ψ_t .

3D-aware rendering and compositing. From a source frame and its tracked parameters Θ_t , GUAVA [31] reconstructs a 3D Gaussian avatar $\mathcal{G}$; we write $\tilde{\mathcal{G}}_t$ for this avatar animated by the motion Θ_t^{mot} at frame t . GUAVA’s Gaussian renderer $\mathcal{R}$ rasterizes $\tilde{\mathcal{G}}_t$ under a camera preset $\mathbf{c}^{(k)}$ and alpha-composites it with the current background stream:

$$\hat{v}_t^{(k)} = \mathcal{R}(\tilde{\mathcal{G}}_t, \mathbf{c}^{(k)}) \odot \alpha_t^{\text{rend},(k)} + b_t \odot (1 - \alpha_t^{\text{rend},(k)}), \quad (2)$$

where $\odot$ is the per-pixel product, $\alpha_t^{\text{rend},(k)}$ is the rendered alpha matte, and b_t is the background frame from the current stream (either v_t or v_t^{bg}). This produces the novel-view variants in Fig. 2a–d.

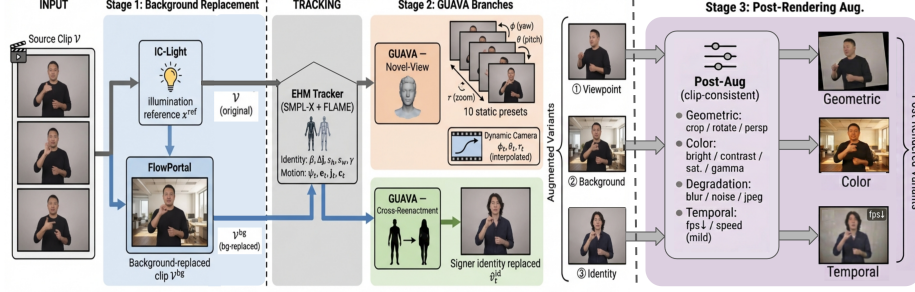


Fig. 1: Overview of our augmentation framework. Stage 1: background replacement (FlowPortal [6] + IC-Light [33]). Stage 2: EHM-Tracker [30] tracks each stream to SMPL-X/FLAME parameters, then GUAVA [31] performs novel-view rendering (static/dynamic cameras) and cross-identity reenactment. Stage 3: post-rendering augmentations (video-consistent image-space transforms and mild temporal resampling).

Viewpoint severity presets. We define $K=10$ spherical-camera presets (ϕ, ϑ, r) (azimuth, elevation, radius) with LookAt = $[0, 0.75, 0]$, indexed by increasing perturbation magnitude (Tab. 2). For each level $\ell \in \{1, \dots, 10\}$, we sample (ϕ, ϑ, r) uniformly within the corresponding ranges. We additionally generate a dynamic-camera trajectory by linearly interpolating between two endpoints sampled from the same level (Fig. 2d), enabling temporally varying viewpoint shifts. The camera orbit space visualization is shown in panel (*) of Fig. 2.

3.2 Background Replacement

We estimate a temporally consistent signer (foreground) soft alpha matte $\alpha_t^{\text{fg}} \in [0, 1]$ via video matting. Guided by α_t^{fg} , we apply illumination-aware background replacement with IC-Light [33] and FlowPortal [6], confining edits to the background region and relighting the scene while preserving the signer’s motion and geometry (Fig. 2f); this produces the background-replaced frame v_t^{bg} .

Lighting severity. Since illumination changes can affect sign-language perception, we stratify background-edited clips into 10 severity levels (L1–L10) using a signer-region photometric shift score (original vs. edited). The exact score definition is provided in Appendix A, and the bin ranges are summarized in Tab. 2.

3.3 Cross-Identity Reenactment

Using the factorization in Eq. (1), we perform cross-identity reenactment by pairing the identity block from one subject with the motion block from another clip. Specifically, we take Θ^{id} from a source subject and Θ_t^{mot} from a target clip to form $\hat{\Theta}_t = (\Theta_{\text{src}}^{\text{id}}, \Theta_{t,\text{tgt}}^{\text{mot}})$, then render the resulting avatar with the target camera/background stream using the same compositor as in Eq. (2). This yields the cross-identity variants shown in Fig. 2e.

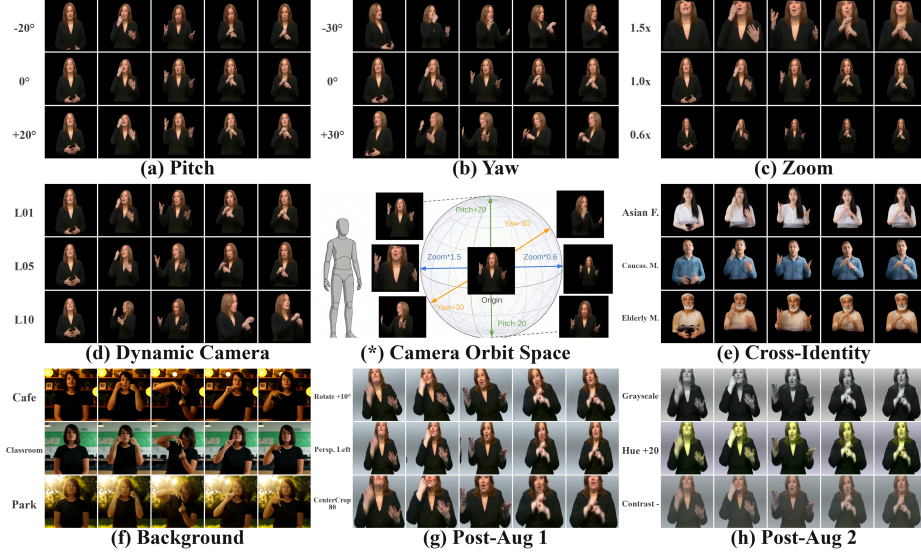


Fig. 2: Qualitative augmentation examples. (*) Camera orbit space visualization. (a–d) Novel-view camera augmentations (pitch, yaw, zoom, dynamic). (e) Cross-identity reenactment. (f) Background replacement. (g–h) Post-rendering augmentations (geometric/photometric/temporal).

3.4 Post-Rendering Augmentations

Finally, we apply lightweight *post-rendering* augmentations to augmented clips. These broaden appearance diversity while preserving linguistic annotations (Fig. 2g–h). We use video-consistent image-space transforms (e.g., crop/rotate, color/contrast jitter, mild blur/compression) and mild temporal resampling (e.g., small speed jitter or frame dropping), which are inexpensive yet effective for increasing low-level variability without changing the underlying sign content.

4 SignNet-1M Dataset

4.1 Basic Information

Data sources. SignNet-1M comprises $\sim 1\text{M}$ **augmented** sign-language clips spanning three languages: DGS, ASL, and CSL, built from seven public or self-collected corpora to diversify lexicon and recording conditions.

The DGS subset ($\sim 100\text{K}$) uses Phoenix14T [1]. The ASL subset ($\sim 700\text{K}$) combines How2Sign [5], OpenASL [20], ASL50K [21] (commercially licensed; our augmented clips in SignNet-1M can be distributed for non-commercial research use), and YouTube-ASL [23]. The CSL subset ($\sim 200\text{K}$) includes CSL-Daily [34] and CSL-News [13]. Fig. 3 summarizes source and augmentation breakdowns.

Augmentation composition. We apply a generative augmentation *framework* to each source clip: (i) 3DGS novel-view synthesis (80%) with

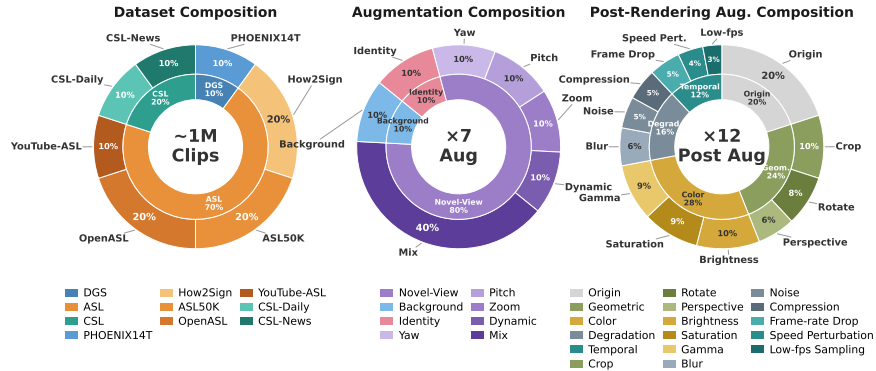


Fig. 3: SignNet-1M composition. **Left:** source-language breakdown (ASL 70%, CSL 20%, DGS 10%). **Middle:** generative augmentation mix (×7): novel-view (80%; yaw/pitch/zoom/dynamic/mix) plus background (10%) and identity (10%) editing. **Right:** post-rendering augmentation mix (×12): origin (20%) plus geometric/color/degradation/temporal augmentations.

yaw/pitch/zoom/ dynamic/mixed augmentations (10/10/10/10/40%), and (ii) diffusion-based appearance editing (20%) split into background (10%) and identity (10%). We further apply lightweight *post-rendering* augmentations (video-consistent image-space transforms and mild temporal resampling) to broaden diversity without changing linguistic annotations. Details are in Sec. 3. Per-language counts and the per-clip computational cost of each augmentation stage are reported in Appendix A.

Data access and license. We provide SignNet-1M resources at <https://signnet.chatsign.ai/>, including (i) train/val/test splits (8:1:1, split at the source-clip level to prevent leakage), (ii) augmentation metadata (factor axes and severity levels), and (iii) reference code and per-stage configurations for the released augmentation components and benchmark checks. The released configs and code support reproducing and extending the documented augmentation components. We distribute the released data under a non-commercial license and provide a standardized data card describing sources, intended use, and known limitations.

4.2 Preprocessing and Annotation

Preprocessing. For YouTube-ASL and OpenASL, we segment clips using sentence-level timestamps [20, 23]. We apply confidence-based pose QC with EHM-Tracker and DWPose [26, 30]: keypoints with confidence < 0.3 are treated as invisible; frames are discarded if either hand has mean confidence < 0.7 or if severe self-occlusion is detected (two-hand distance $MSE < 3072$).

Annotation. All linguistic annotations (gloss/translation/timestamps) are inherited from the source corpora unchanged.

Table 1: Comparison with existing sign language translation datasets. *Aug.*: synthetic augmentation; *Ann.*: G = gloss, T = translation; T (+G) indicates translations throughout and glosses only for Phoenix14T/CSL-Daily subsets.

Dataset	Lang.	Clips	Dur. (h)	Signers	Aug.	Ann.
Phoenix14T [1]	DGS	8.26K	11	9	✗	G+T
How2Sign [5]	ASL	35.00K	80	11	✗	T
OpenASL [20]	ASL	98.42K	288	~220	✗	T
YouTube-ASL [23]	ASL	610.19K	984	~2.5K	✗	T
CSL-Daily [34]	CSL	20.65K	23	~10	✗	G+T
CSL-News [13]	CSL	751.32K	1,985	~20	✗	T
SignNet-1M (ours)	DGS/ASL/CSL	1M	2,058	~10K ⁴	✓	T (+G)

4.3 Dataset Statistics

Tab. 1 compares SignNet-1M with prior translation datasets. The count-weighted average clip length is 7.41 s; per-dataset statistics are provided in Appendix A.

5 Benchmark Suite and Augmented Evaluation Protocol

5.1 Downstream Tasks

We evaluate on two complementary sign language understanding tasks.

Sign Language Translation (SLT) maps a video clip to a spoken-language sentence and is the primary task of our benchmark. We report BLEU-4 [17] ($\uparrow$) as the main metric throughout the paper; ROUGE-L [14] and ChrF [19] results are provided in the supplementary material. All included source corpora provide sentence-level translations, so SLT experiments cover the full SignNet-1M benchmark across DGS, ASL, and CSL.

Sign Language Recognition (SLR) maps a video clip to a gloss sequence and is evaluated by Word Error Rate (WER%, $\downarrow$). Because gloss annotations are only available in Phoenix14T [1] and CSL-Daily [34], SLR experiments are restricted to the DGS and CSL sub-datasets of SignNet-1M.

5.2 Evaluation Protocol

Each task is evaluated under four settings to measure both robustness gaps in existing data and the training value of SignNet-1M:

- **Setting A:** train and evaluate on the source benchmark.

⁴ The reported number of signers in SignNet-1M includes *synthetic identities* introduced by cross-identity reenactment, in addition to original signers from the source corpora.

- **Setting B:** train on the source benchmark and evaluate on the SignNet-1M test split.
- **Setting C:** train and evaluate on SignNet-1M.
- **Setting D:** train on SignNet-1M and evaluate on the source benchmark test set, measuring whether augmented training helps under the natural source distribution.

For compactness, result tables refer to these settings simply as A–D. The SignNet-1M test split is obtained by augmenting the *source test set*; splits are made at the source-clip level, so no train/test leakage occurs between Settings. We define **Gap** = B – A and **Gain** = C – B (higher is better); for lower-is-better metrics (e.g., WER%), we flip the sign. Negative Gap indicates sensitivity to viewpoint/background/identity/post-rendering shifts, while positive Gain shows that SignNet-1M improves robustness under such shifts.

5.3 Severity-Stratified Evaluation Factors

Beyond aggregate performance, we stratify the SignNet-1M test set by augmentation severity to reveal how robustness degrades as visual conditions become more challenging. Severity levels are defined along a viewpoint axis (static & dynamic) and a scene-lighting axis, and summarised in Tab. 2.

Table 2: Ranges of parameters and severity-level definitions for stratified evaluation.

Axis	Levels	Definition
Viewpoint (static)	L1–L10	$ \phi : 0^\circ \rightarrow 30^\circ, \vartheta : 0^\circ \rightarrow 25^\circ, r : [1.0, 1.0] \rightarrow [0.6, 1.5]$
Viewpoint (dynamic)	L1–L10	Sample two endpoints from the corresponding static Lk range and linearly interpolate $(\phi_t, \vartheta_t, r_t)$ over time.
Lighting	L1–L10	$\hat{s}_{\text{light}} \in [0, 1]$: normalized signer-region photometric shift, binned into 10 equal-width levels.

Viewpoint severity. We evaluate static and dynamic camera variants using the $K=10$ severity bins defined in Sec. 3.1, indexed by increasing augmentation magnitude.

Scene severity. For diffusion-based background editing, we define 10 linearly spaced levels along a single axis of lighting variation. Lighting levels are computed from a normalized signer-region photometric severity score, with parameter ranges in Tab. 2.

Both Setting B and Setting C results are reported per level in Sec. 6.3, enabling fine-grained diagnosis of where and how model performance degrades.

Table 3: Sign language translation metric BLEU-4 ($\uparrow$) on sub-datasets of SignNet-1M. Evaluation conditions and Δ metrics are defined in Sec. 5.2. Columns A–D correspond to Settings A–D in Sec. 5.2.

Sub-dataset	Method	Source test	SignNet-1M test		Source test	Δ	
		A	B	C	D	Gap	Gain
Phoenix14T	SpaMo [9]	22.49	7.81	18.98	27.78	-14.68	+11.17
	GASLT [27]	15.74	1.26	14.95	–	-14.48	+13.69
	FPT4SLT [4]	19.18	7.25	18.72	–	-11.93	+11.47
How2Sign	SpaMo [9]	10.11	6.25	10.36	18.48	-3.86	+4.11
	UniSign [13]	14.50	6.47	12.83	–	-8.03	+6.36
OpenASL	SpaMo [9]	10.32	0.34	9.71	–	-9.98	+9.37
	UniSign [13]	22.67	8.12	22.83	–	-14.55	+14.71
CSL-Daily	SpaMo [9]	20.55	12.70	21.32	20.93	-7.85	+8.62
	UniSign [13]	25.27	13.47	21.66	–	-11.80	+8.19
	GASLT [27]	4.07	0.12	3.68	–	-3.95	+3.56

6 Experiments

6.1 Experimental Setup

Datasets. We evaluate on four SignNet-1M sub-datasets spanning DGS, ASL, and CSL: Phoenix14T, How2Sign, OpenASL, and CSL-Daily. All results follow the unified four-setting protocol in Sec. 5.2.

Models. We benchmark representative state-of-the-art backbones from distinct model families. For translation, we include SpaMo [9] and additionally report GASLT [27] and FPT4SLT [4] where applicable. For unified recognition/translation, we evaluate UniSign [13]. For recognition on Phoenix14T, we further include an online CSLR baseline [36].

Training and evaluation. For fair comparison, we follow each method’s official training and evaluation setup and vary only the training/evaluation data source (source benchmark vs. SignNet-1M). We report BLEU-4 for translation and WER% for recognition; additional metrics and all method-specific hyperparameters are reported in the Appendix. We further conduct severity-stratified evaluations (Sec. 6.3), ablations (Sec. 6.4), scaling experiments (Sec. 6.5), and visual-quality checks (Sec. 6.6).

6.2 Translation & Recognition Performance

We evaluate SignNet-1M on sign language translation (SLT) and recognition (SLR) under the unified protocol in Sec. 5.2.

Translation performance. Tab. 3 reports BLEU-4 under Settings A–D. Setting B drops sharply relative to Setting A even for strong baselines (e.g.,

Table 4: Generalization across model families. Δ is the improvement of SignNet-1M-trained over source-benchmark-trained models (BLEU-4 for SLT methods, WER% for the Online-CSLR recognizer). SOTA methods span three typical model-family specifications used in the sign-language community.

Method	Modality	Supervision	Task	Δ
SpaMo [9]	RGB	Gloss-free	SLT	+11.17
FPT4SLT [4]	RGB	Gloss-free	SLT	+11.47
GASLT [27]	RGB	Gloss-free	SLT	+13.69
UniSign [13]	RGB+Pose	Gloss-based	SLT	+14.71
Online-CSLR [36]	RGB+Pose	Gloss-based	SLR	+23.26%

Table 5: Sign language recognition metric WER% ($\downarrow$) on sub-datasets of SignNet-1M. Evaluation conditions and Δ metrics are defined in Sec. 5.2.

		Source test SignNet-1M test			Δ	
Sub-dataset	Method	A	B	C	Gap	Gain
Phoenix14T	Online-CSLR [36]	22.21%	51.45%	28.19%	-29.24%	+23.26%
CSL-Daily	UniSign [13]	28.20%	56.90%	30.74%	-28.70%	+26.16%

Phoenix14T: -11.93 to -14.68 ; OpenASL: -9.98 to -14.55), indicating sensitivity to the factor shifts in SignNet-1M. Setting C consistently recovers performance with large gains over Setting B across languages and architectures (e.g., Phoenix14T: $+11.17$ to $+13.69$; OpenASL: $+9.37$ to $+14.71$), demonstrating improved robustness under challenging visual conditions. Crucially, models trained on SignNet-1M also improve on the source benchmark test sets (Setting D vs. Setting A): SpaMo gains $+5.29$, $+8.37$, and $+0.38$ BLEU-4 on Phoenix14T, How2Sign, and CSL-Daily, showing the augmentation helps under natural distributions, not only constructed shifts. Across three typical model-family specifications (RGB gloss-free, RGB+Pose gloss-based SLT, and pose-based recognition), training on SignNet-1M consistently improves over source-benchmark training (Tab. 4), indicating the gains are not tailored to a single architecture.

Recognition performance. Tab. 5 shows a similar trend for WER% ($\downarrow$). Setting B degrades substantially relative to Setting A on SignNet-1M (Phoenix14T: $22.21\% \rightarrow 51.45\%$; CSL-Daily: $28.20\% \rightarrow 56.90\%$), while Setting C markedly reduces WER (to 28.19% and 30.74% , respectively), yielding large gains ($+23.26\%$ and $+26.16\%$). This provides evidence that SignNet-1M improves robustness for both translation and recognition, rather than only boosting average-case accuracy.

6.3 Experiments Under Increasing Severity

Following the severity stratification in Sec. 5.3, we report BLEU-4 across severity levels (L1–L10) under Settings B and C on SignNet-1M (Phoenix14T

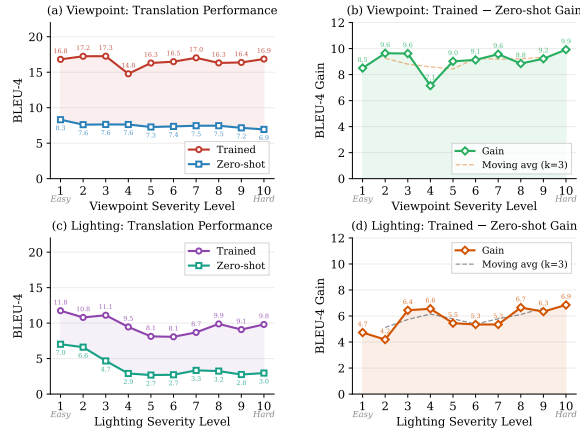


Fig. 4: Stratified SLT performance across severity levels on Phoenix14T. (a,c) BLEU-4 of Settings B and C. (b,d) Gain (C–B) in BLEU-4. Viewpoint levels (L1–L10) aggregate static yaw/pitch/zoom and dynamic-camera variants within each severity bin, and lighting levels (L1–L10) are defined by a normalized photometric shift defined in Tab. 2.

sub-dataset, SpaMo) [9]. Fig. 4 summarizes robustness under increasing viewpoint augmentations and lighting shifts.

Viewpoint. As viewpoint severity increases (Fig. 4a), Setting B BLEU-4 steadily degrades ($8.31 \rightarrow 6.94$), while Setting C yields higher scores with a much flatter drop. Accordingly, the training gain grows at harder levels (Fig. 4b; $8.50 \rightarrow 9.92$), indicating improved robustness under severe viewpoint shifts.

Lighting. Lighting shifts induce a larger Setting B collapse (Fig. 4c; $7.02 \rightarrow 2.96$), whereas Setting C substantially mitigates the drop and remains consistently stronger across levels. The gain also increases with severity (Fig. 4d; $4.73 \rightarrow 6.85$), suggesting training is particularly effective under extreme illumination changes. Overall, larger gains at harder levels show that covering such shifts is essential for in-the-wild robustness, motivating SignNet-1M.

6.4 Ablation Study

Tab. 6 analyzes the contribution of individual augmentation components in SignNet-1M on Phoenix14T and How2Sign under Settings B and C. Among the novel-view factors, **zoom** yields the most consistent gains, while **dynamic camera** is less stable, indicating a harder shift that benefits from careful severity control. For appearance variation, *identity* editing yields larger gains than *scene* editing in the tested configurations (e.g., 13.06/22.04 vs. 4.91/12.53 on Phoenix14T), highlighting signer diversity as an important robustness source. *Post-rendering augmentation* adds another source of low-level variation on top of rendering-based augmentation. Overall, different augmentation factors contribute under different shifts; the full SignNet-1M mix is designed for broad factor coverage rather than optimizing a single isolated subset metric.

Table 6: Ablation of augmentation components (BLEU-4). Results are reported on sub-datasets of SignNet-1M (Phoenix14T and How2Sign) under Settings B and C using SpaMo method [9].

Configuration	Phoenix14T		How2Sign	
	B	C	B	C
<i>Novel View</i>	7.10	18.85	6.09	10.32
+ yaw	8.17	18.59	7.09	10.78
+ pitch	6.98	18.36	4.06	9.04
+ zoom	8.89	18.98	7.81	11.05
+ dynamic camera	6.87	17.13	6.16	8.70
-----	-----	-----	-----	-----
<i>Scene editing</i>	4.91	12.53	3.28	9.26
-----	-----	-----	-----	-----
<i>Identity editing</i>	13.06	22.04	8.67	12.54
-----	-----	-----	-----	-----
<i>Post-rendering aug.</i>	5.37	18.16	2.51	10.17
-----	-----	-----	-----	-----
<i>Full SignNet-1M</i>	7.81	18.98	6.25	10.36

Table 7: Visual quality comparison between original and SignNet-augmented videos. Lower FID-VID/FVD and higher SSIM/PSNR indicate better quality.

Sub-Dataset	FID-VID↓		FVD↓		SSIM↑	PSNR↑
	Origin	SignNet-Aug	Origin	SignNet-Aug		
Phoenix14T [1]	49.96	49.22	141.57	119.82	0.92	25.30
How2Sign [5]	47.26	48.68	118.69	123.84	0.91	27.98
ASL50K [21]	43.41	44.06	119.15	137.52	0.83	21.83
OpenASL [20]	49.52	48.29	157.79	121.44	0.81	23.00
CSL-Daily [34]	40.39	39.80	124.45	136.34	0.88	23.24

6.5 Scaling Experiments

We study how translation performance scales with the amount of augmented data by training with increasing scale factors K ($K=2, 5, 10$; Fig. 5), where K is the augmented-set size relative to the original data ($K=10$ is a $10\times$ expansion). Holding the backbone (SpaMo [9]) and training epochs fixed (Sec. 5.2), BLEU-4 increases monotonically with K on both Phoenix14T and How2Sign [1, 5] ($14.37 \rightarrow 18.54$ and $10.23 \rightarrow 14.11$), indicating that larger augmentation budgets yield stronger robustness, motivating the scaling of SignNet-1M to million-level volume. Matched-GPU-hour comparisons are provided in the Appendix.

6.6 Visual Quality Assessment

Qualitative analysis. Fig. 2 shows representative outputs: novel-view renderings preserve hand/face structure under camera shifts; cross-identity reenactment maintains motion synchrony; background replacement changes scene

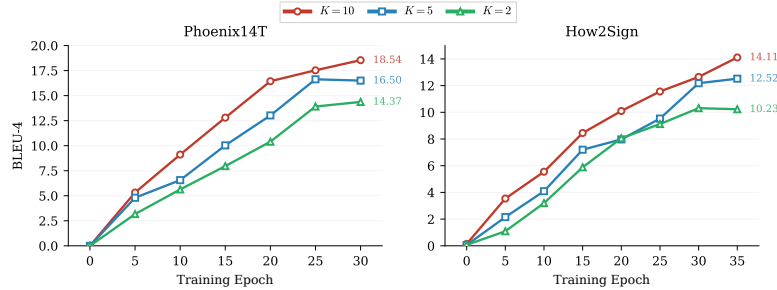


Fig. 5: Scaling with augmentation factor K . BLEU-4 vs. training epoch when training with increasing augmentation scale factors ($K=2, 5, 10$). Results are shown on Phoenix14T (left) and How2Sign (right).

Table 8: Identity-swap fidelity. FID-VID↓ for identity-swapped vs. non-identity-swapped clips, and BLEU-4 of models trained with vs. without identity editing. Identity swapping changes FID-VID by only $\sim 2\%$ and keeps translation quality comparable.

Sub-dataset	FID-VID↓		BLEU-4↑	
	w/o ID swap	w/ ID swap	w/o ID edit	w/ ID edit
Phoenix14T [1]	50.99	49.67	22.49	22.04
How2Sign [5]	47.26	48.05	10.11	12.54
CSL-Daily [34]	40.39	39.61	—	—

context; and post-rendering augmentations simulate common capture degradations while largely preserving linguistic content.

Quantitative analysis. We assess visual quality with distribution-level FID-VID [8]/FVD [22] and pairwise PSNR/SSIM [24] (Tab. 7). SignNet-Aug stays close to the originals: relative to Origin, FID-VID changes within $[-2.5\%, +3.0\%]$, FVD within $[-23.0\%, +15.4\%]$, and SSIM remains high (≥ 0.81), indicating strong distributional realism and structural consistency.

Identity-swap fidelity. We compare identity-swapped and non-identity-swapped clips (Tab. 8): FID-VID shifts by only $\sim 2\%$ and translation remains comparable (e.g., How2Sign BLEU-4 $10.11 \rightarrow 12.54$), suggesting that identity swaps largely preserve articulation and temporal consistency.

6.7 Human Evaluation

We conducted two IRB-approved studies with 14 hired native ASL signers (Tab. 9). On 27,000 source/synthesized ASL sentence pairs from How2Sign [5], OpenASL [20], and ASL50K [21], novel-view, background, and post-rendering augmentations reached $> 99\%$ acceptance, and cross-identity reenactment reached 94.8%. On a self-collected set of 175 ASL sentences and 500 isolated

Table 9: Human evaluation by 14 native ASL signers. Acceptance = fraction of synthesized clips judged to preserve the source sign and its linguistic content.

Study / augmentation	#Samples	Acceptance
Novel-view / background / post-rendering	27,000 pairs	> 99%
Cross-identity reenactment	included in 27,000 pairs	94.8%
Self-collected (175 sentences, 500 words)	675	100%

words, acceptance was 100%. These results support semantic preservation on the evaluated ASL subsets.

7 Limitations

(1) Since GUAVA [31] reconstructs avatars from a single image, rendering quality may degrade under large viewpoint changes, especially for fine hand details and loose clothing. (2) SignNet-1M covers DGS/ASL/CSL, so results may not fully generalize to other languages. (3) Identity images for cross-reenactment come only from publicly released corpora with research consent; synthesized clips are metadata-tagged to prevent re-identification misuse. (4) The benchmark covers representative model families, but the number and diversity of evaluated architectures remain limited.

8 Conclusion

We presented SignNet-1M, a multilingual augmented sign language dataset with ~ 1 M clips across ASL, DGS, and CSL. By combining 3DGS novel-view rendering, diffusion-based scene/identity editing, and lightweight post-rendering augmentations, SignNet-1M creates controlled shifts in viewpoint, background, and signer appearance while preserving sign motion and annotations. Our four-setting evaluation reveals substantial robustness gaps and shows that SignNet-1M improves robustness for real-world deployment. We provide the dataset, released augmentation-component code/configs, metadata, and benchmark resources at <https://signnet.chatsign.ai/>.

Acknowledgements

This work was partially supported by ChatSign Technology, Ltd., the NYUAD Center for Interdisciplinary Data Science & AI (CIDSAI), funded by Tamkeen under the NYUAD Research Institute Award CG016, and the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001.

References

1. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: CVPR. pp. 7784–7793 (2018). <https://doi.org/10.1109/CVPR.2018.00812>
2. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: CVPR. pp. 10023–10033 (2020). <https://doi.org/10.1109/CVPR42600.2020.01004>
3. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: Practical automated data augmentation with a reduced search space. In: CVPRW. pp. 3008–3017 (2020). <https://doi.org/10.1109/CVPRW50498.2020.00359>
4. De Coster, M., D’Oosterlinck, K., Pizurica, M., Rabaey, P., Verlinden, S., Van Herreweghe, M., Dambre, J.: Frozen pretrained transformers for neural sign language translation. In: 1st International Workshop on Automated Translation for Signed and Spoken Languages (August 2021)
5. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: A large-scale multimodal dataset for continuous american sign language. In: CVPR. pp. 2735–2744 (2021)
6. Gao, W., Fan, J., Zeng, J., Yang, S.: FlowPortal: Residual-corrected flow for training-free video relighting and background replacement. arXiv preprint arXiv:2511.18346 (2025)
7. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: AugMix: A simple data processing method to improve robustness and uncertainty. In: ICLR (2020)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
9. Hwang, E.J., Cho, S., Lee, J., Park, J.C.: An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In: Proc. Nations Am. Chapter Assoc. Comput. Linguistics. pp. 3901–3920 (2025). <https://doi.org/10.18653/v1/2025.naacl-long.197>
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4), 139:1–139:14 (2023). <https://doi.org/10.1145/3592433>
11. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: ICCV. pp. 15954–15964 (2023)
12. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM TOG **36**(6), 194:1–194:17 (2017). <https://doi.org/10.1145/3130800.3130813>
13. Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-sign: Toward unified sign language understanding at scale. In: ICLR (2025)
14. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, Proc. ACL Workshop. pp. 74–81 (2004)
15. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: CVPR. pp. 8599–8608 (2024)
16. Mukushev, M., Ubingazhibov, A., Kydyrbekova, A., Imashev, A., Kimmelman, V., et al.: Fluentsigners-50: A signer independent benchmark dataset for sign language processing. PLOS ONE **17**(9), e0273649 (2022). <https://doi.org/10.1371/journal.pone.0273649>

17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: *Proc. ACL*. pp. 311–318 (2002)
18. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: *CVPR*. pp. 10975–10985 (2019). <https://doi.org/10.1109/CVPR.2019.01123>
19. Popović, M.: chrF: Character n-gram F-score for automatic MT evaluation. In: *Proc. Workshop on Statistical Machine Translation (WMT)*. pp. 392–395 (2015)
20. Shi, B., Brentari, D., Shakhnarovich, G., Livescu, K.: Open-domain sign language translation learned from online video. In: *Proc. Conf. Empirical Methods Nat. Lang. Process.* pp. 6365–6379 (2022). <https://doi.org/10.18653/v1/2022.emnlp-main.427>
21. SignASL.org: SignASL: American sign language video dictionary. <https://www.signasl.org> (2024), commercially licensed ASL video corpus; the ASL50K subset used in this work was obtained under license. Accessed 28 Jun 2026
22. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinière, R., Michalski, M., Gelly, S.: FVD: A new metric for video generation. *arXiv preprint arXiv:1812.01717* (2019)
23. Uthus, D., Tanzer, G., Georg, M.: YouTube-ASL: A large-scale, open-domain American Sign Language–English parallel corpus. In: *NeurIPS* (2023)
24. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE TIP* **13**(4), 600–612 (2004). <https://doi.org/10.1109/TIP.2003.819861>
25. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: *ICCV*. pp. 7623–7633 (2023)
26. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. pp. 4210–4220 (2023), https://openaccess.thecvf.com/content/ICCV2023W/CV4Metaverse/papers/Yang_Effective_Whole-Body_Pose_Estimation_with_Two-Stages_Distillation_ICCVW_2023_paper.pdf, accessed 28 Jun 2026
27. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., Zhao, Z.: Gloss attention for gloss-free sign language translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2551–2562 (2023)
28. Yin, K., Read, J.: Better sign language translation with STMC-transformer. In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. pp. 5975–5989 (2020). <https://doi.org/10.18653/v1/2020.coling-main.525>, <https://aclanthology.org/2020.coling-main.525/>, accessed 28 Jun 2026
29. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: Regularization strategy to train strong classifiers with localizable features. In: *ICCV*. pp. 6023–6032 (2019). <https://doi.org/10.1109/ICCV.2019.00612>
30. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: EHM-Tracker: Official ehm tracking implementation for GUAVA. <https://github.com/Pixel-Talk/EHM-Tracker> (2025), software release. Accessed 28 Jun 2026
31. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: Guava: Generalizable upper body 3d gaussian avatar (2025)
32. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: *ICLR* (2018)

33. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: ICLR (2025)
34. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: CVPR. pp. 1316–1325 (2021)
35. Zou, Y., Choi, J., Wang, Q., Huang, J.: Learning representational invariances for data-efficient action recognition. CVIU **227**, 103597 (2023). <https://doi.org/10.1016/j.cviu.2022.103597>
36. Zuo, R., Wei, F., Mak, B.: Towards online continuous sign language recognition and translation. In: EMNLP (2024)