SIGNXRAY: PLUG-AND-PLAY SIGN ENCODING IN A FACE-HAND-BODY MULTI-PRIOR SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

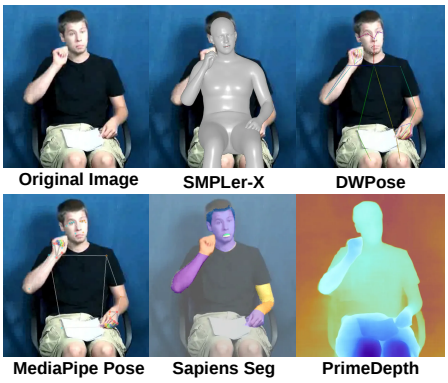
Sign languages convey information through coordinated facial expressions, hand articulation, body motion, and their temporal interactions. Existing human-representation models expose complementary aspects of this signal through heterogeneous outputs that are difficult to reuse through a common video interface. We present **SignXray**, a plug-and-play sign video encoder learned in a unified Face-Hand-Body multi-prior space. During training, SignXray consolidates supervision from SMPLer-X, DWPose, MediaPipe, PrimeDepth, and Sapiens Segmentation into a compact, sign-aware representation and distills this heterogeneous prior knowledge into an RGB ViT. At inference time, the resulting encoder maps RGB video directly into the unified representation, giving downstream models access to the complementary information of all five prior formats through a single reusable feature interface. Experimental results demonstrate that SignXray achieves state-of-the-art performance on continuous SLR and translation tasks, with up to a 50$\times$ acceleration over the evaluated pixel-space baseline.

1 INTRODUCTION

Sign Language Understanding (SLU) includes recognition and translation tasks that map sign-language videos to gloss sequences or written-language text. These technologies have important social value in facilitating barrier-free communication between deaf and hearing individuals Camgöz et al. (2018); Cihan Camgöz et al. (2020); Stoll et al. (2020); Yin et al. (2021). SLU presents two closely related representation challenges. First, signed languages convey information through coordinated hand and arm movements, facial expressions, and body postures, with complex temporal dependencies among these channels Boháček & Hrúz (2022); Vintar et al. (2012); Von Agris & Kraiss (2010). Previous work has combined RGB and keypoint streams or fused multiple spatial–temporal cues for recognition and translation Chen et al. (2022b); Zhou et al. (2021c). Recent representation-learning methods further integrate face, hand, body, and RGB streams through unified or self-supervised objectives Li et al. (2025); Gueuwou et al. (2025b;a). These studies establish the value of complementary visual evidence for sign understanding. As illustrated in Fig. 1, human-representation models expose a broader collection of such evidence through different output structures, dimensionalities, and coordinate conventions. Consolidating 2D and 3D pose, depth, and segmentation cues into a single reusable video representation therefore remains non-trivial.

To address these challenges, we develop **SignXray**, a plug-and-play RGB video encoder that compiles heterogeneous human-representation priors into a single sign-specific feature interface. Our central insight is that five complementary human-representation models, spanning pose, depth, and segmentation, can jointly specify the visual information a sign encoder should preserve. SignXray converts this complementary expert knowledge into a reusable representation that is produced directly from RGB video. The five priors describe complementary aspects of signing: SMPLer-X provides parametric 3D body structure; DWPose captures 2D body, hand, and dense facial landmarks; MediaPipe supplies lightweight 3D joints; PrimeDepth encodes front–back spatial relationships; and Sapiens captures part-level body shape. Their outputs differ substantially in topology, dimensionality, and granularity. SignXray preserves each source through a type-specific encoder, integrates the resulting sequences with attention, and projects them into a shared 2,048-dimensional Face-Hand-Body space. Sign-semantic alignment organizes this space around signed content, while variance-based dimension masking concentrates information in its active channels Saunders et al. (2020b); Stoll et al. (2020); Moryossef (2022). The unified space then serves as a frozen repre-

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.



Multimodal pose estimation methods. SMPLer-X Cai et al. (2023) provides accurate 3D human body parameters; DWPose Yang et al. (2023) focuses on real-time 2D keypoint detection; MediaPipe Lugaresi et al. (2019) provides lightweight and efficient 3D pose prediction; PrimeDepth Zavadski et al. (2024) estimates scene depth; and Sapiens Segmentation Khirodkar et al. (2024) provides fine-grained human body part segmentation. SignXray regards them as heterogeneous, complementary human-centric priors. Their consolidation into a unified Face-Hand-Body space produces information-dense video representations for diverse downstream sign-understanding models. Inference uses the distilled RGB representation directly.

Figure 1: **Unified Face-Hand-Body multi-prior space.** SignXray integrates complementary body, hand, face, depth, and segmentation cues from five heterogeneous human-representation sources.

sentation target for RGB video learning. A spatiotemporal ViT predicts the same feature sequence directly from video frames, compiling the five expert pipelines into the parameters of one encoder. At inference time, SignXray exposes a fixed RGB-to-feature interface that connects to recognition and translation heads. This separation between representation learning and task decoding enables direct encoder substitution across downstream architectures Dosovitskiy et al. (2021); Boháček & Hrúz (2022); Cho et al. (2021); Chiang et al. (2019).

This multi-prior compilation is the central mechanism of SignXray. We evaluate the resulting encoder on ASL and mainstream benchmarks spanning continuous recognition, isolated recognition, and translation Neidle & Metaxas (2025); Forster et al. (2012); Zhou et al. (2021a); Li et al. (2020a).

The main contributions of this paper can be summarized in the following three points:

- We construct a **unified Face-Hand-Body multi-prior space** that consolidates complementary body, hand, face, depth, and segmentation cues into an information-dense sign representation.
- We develop **SignXray**, a ViT-based RGB video encoder distilled from this heterogeneous supervision. Its single reusable RGB-to-feature interface gives downstream models direct access to the complementary knowledge of all five sources.
- We establish a **plug-and-play interface across sign-language models and tasks**. Its fixed RGB-to-feature output supports direct encoder substitution in downstream architectures. Encoder-adapted heads evaluate the representation on continuous recognition, isolated recognition, and translation.

2 Related Work

2.1 Sign Language Recognition

Sign language recognition is an important research direction at the intersection of computer vision and natural language processing. Early works mainly relied on hand-crafted features and rules to recognize small sign language vocabularies Grobel & Assan (1997); Kadir et al. (2004). Subsequently, methods based on machine learning, including hidden Markov models and 3D human pose and motion analysis, were used to recognize ASL Vogler & Metaxas (1998); Metaxas (1999). With the development of deep learning, neural network-based methods began to dominate this field. Early deep approaches used CNNs Hershey et al. (2017) to extract spatial features from video frames and RNNs Sutskever et al. (2014a) to model temporal relationships Cui et al. (2017); Koller et al. (2015). These systems established the spatial–temporal recognition pipeline. Subsequent work increasingly focused on representing fine-grained signing dynamics Fang et al. (2025b).

Researchers have also introduced human pose estimation Katz (1987); Ko et al. (2019a); Charles et al. (2014); Cihan Camgöz et al. (2020); Moryossef et al. (2020); Yin et al. (2022), converting sign-language videos into skeleton sequences and improving the modeling of detailed motion Saunders

et al. (2020b; 2021a). As increasingly capable human-representation models have emerged Cai et al. (2023); Yang et al. (2023), their incompatible formats and different representational granularity have made heterogeneous integration an important problem Stoll et al. (2020); Huang et al. (2021); Fang et al. (2025a;c). Signed-language understanding further requires interactions between manual motion and non-manual signals, which can play grammatical roles.

2.2 Sign Language Representation Learning

Recent work has developed transferable sign representations across tasks and visual streams. Uni-Sign Li et al. (2025) learns unified representations at scale by integrating multiple sign-related visual cues. SHuBERT Gueuwou et al. (2025b) learns contextual representations through self-supervised prediction over clustered hand, face, and body streams. SignMusketeers Gueuwou et al. (2025a) uses efficient frame-level face, hand, and body streams for large-scale sign language translation, while SignRep Wong et al. (2025) develops efficient and scalable self-supervised sign representations. These studies demonstrate the value of multi-stream supervision and transferable sign features. Recent translation-specific work aligns RGB video features with text within a dedicated model: VSLT uses a variational prior–posterior objective, and SignLlama adapts visual features for LLM decoding Zhao et al. (2026); Gan et al. (2026a). Contemporaneous methods further explore explicit multi-granularity video–text alignment (VTaMo), latent semantic reasoning (SignThought), motion-level expert transfer across languages (SIGNET), diffusion-based semantic refinement (BoostSLT), and multilingual discrete units (Q-BridgeNet) Hu et al. (2026); Jiang et al. (2026); Asasi et al. (2026); Han et al. (2026); Feng et al. (2026).

2.3 Pose and Human Representation Estimation

Pose estimation technology has made significant progress in recent years. SMPLer-X Cai et al. (2023) models the body, hands, and face through complete 3D human-body parameter estimation Loper et al. (2015); Cao et al. (2017). DWPose Yang et al. (2023) focuses on stable, real-time 2D keypoint detection Wang et al. (2018a); Lugaresi et al. (2019). MediaPipe Lugaresi et al. (2019) provides lightweight 3D pose prediction for real-time applications Chan et al. (2019).

Beyond joint coordinates, complementary human-centric estimators expose additional geometric and appearance cues. PrimeDepth Zavadski et al. (2024) provides depth and 3D structural information Qiu et al. (2021); Carreira & Zisserman (2017). Sapiens segmentation Khirodkar et al. (2024) captures fine-grained human-part boundaries Güler et al. (2018); Saunders et al. (2021b). Their complementary representations for sign-language recognition Szegedy et al. (2015); Koller (2020) span substantially different dimensionalities and structures. SignXray uses them as heterogeneous training-time priors and distills their knowledge into a common RGB video encoder.

3 SignXray Multi-Prior Sign Encoder

3.1 Data Processing

To address the challenge of integrating heterogeneous pose information for SL recognition, we first construct a comprehensive processing pipeline that standardizes multiple pose formats. First, we use a high-quality ASLLRP SignStream® 3 Corpus Neidle et al. (2022a); Neidle & Metaxas (2025) dataset, which contains over 80 hours of American Sign Language (ASL) videos with synchronized front view, side view, and facial close-up recordings. For each input video $V \in \mathbb{R}^{T\times H\times W\times 3}$, where T represents the number of frames and H, W represent height and width, respectively, we extract sequential pose representations for each pose type: $P_{\text{DWPose}} \in \mathbb{R}^{T\times 384}$, $P_{\text{MediaPipe}} \in \mathbb{R}^{T\times 258}$, $P_{\text{SMPLer-X}} \in \mathbb{R}^{T\times 165}$, $P_{\text{PrimeDepth}} \in \mathbb{R}^{T\times 576}$, $P_{\text{Sapiens}} \in \mathbb{R}^{T\times 576}$. The dimensionality of each pose type reflects its specific information structure. For instance, DWPose contains 18 body keypoints, 21 keypoints for each hand, and 68 facial keypoints, along with their respective confidence scores, totaling 384 dimensions per frame.

As shown in Fig. 2, we flatten each per-frame output and concatenate the five pose types into a 1,959-dimensional record. Type-specific encoders preserve the identity of each source, multi-head attention integrates their complementary information, and a spatial matching projection maps the

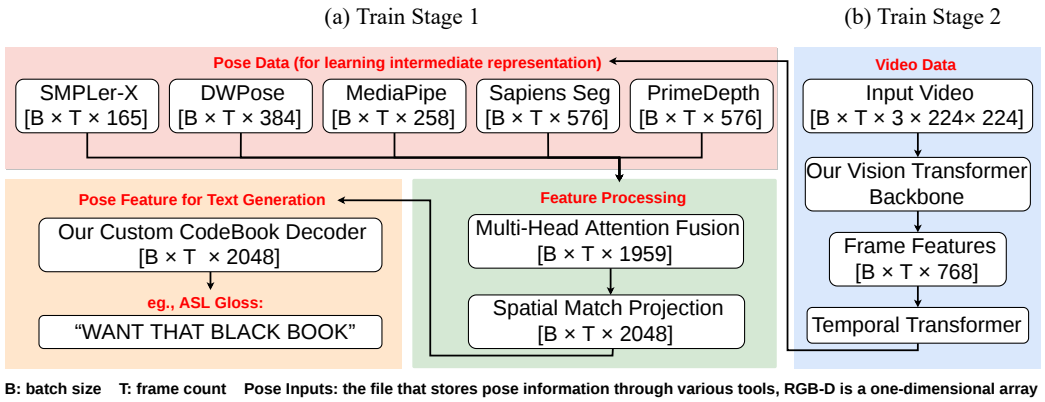


Figure 2: **Learning the SignXray multi-prior encoder.** (a) In Stage 1, heterogeneous face, hand, body, depth, and segmentation priors supervise a unified sign-aware representation. (b) In Stage 2, the prior-space teacher is frozen and an RGB ViT learns to reproduce its information-dense features. At inference time, the resulting RGB encoder directly provides the distilled multi-prior knowledge to different sign-understanding heads.

fused feature to the 2,048-dimensional unified space. This standardized structure isolates type-specific signals and establishes compatibility among the heterogeneous pose formats.

3.2 Multimodal Pose Fusion

We construct the unified space with auxiliary sign-semantic supervision, so the fused pose features encode information relevant to language understanding. The Stage-1 model contains type-specific pose encoders, a multimodal fusion layer, ViT feature-processing layers, and a lightweight text decoder used during training.

We choose ViT for pose latent space construction because: (1) its self-attention mechanism fuses heterogeneous pose representations from SMPLer-X, DWPose, MediaPipe, PrimeDepth, and Sapiens; (2) its deep layered architecture creates a hierarchical latent space that compresses rich multi-source information and preserves essential pose details; and (3) its transformer architecture scales effectively to high-dimensional inputs He et al. (2022). These properties support a compact, pose-rich space for sign-language video encoding.

Multimodal Pose Fusion Layer. The pose encoder E_{pose} first processes the five pose types with source-specific encoders and concatenates their outputs along the feature dimension:

$$\begin{aligned} f_i &= E_i(P_i), \\ F &= \text{Concat}(f_1, f_2, \ldots, f_5), \\ f_{\text{fused}} &= \text{MultiHeadAttention}(F). \end{aligned} \tag{1}$$

Here, E_i is a width-preserving source adapter for pose type i, f_i is its encoded sequence with the same per-frame width as P_i, and $F \in \mathbb{R}^{T\times 1959}$ is the concatenated multi-prior representation. This fusion preserves source-specific characteristics and enables information sharing across the five priors Ko et al. (2019b); Liu et al. (2022); Zhang et al. (2020).

We then map the fused feature to the common 2,048-dimensional interface and process it with the ViT layers:

$$\begin{aligned} z^{(0)} &= \text{PadMatch}(f_{\text{fused}}) \in \mathbb{R}^{T\times 2048}, \\ Z_{\text{pose}} &= \text{ViT}_{\text{layers}}(z^{(0)}). \end{aligned} \tag{2}$$

Here, PadMatch denotes zero-padding followed by a learned projection, and Z_{pose} is the unified pose-rich target representation Stoll et al. (2020); Cho et al. (2021); Lu et al. (2023a).

Sign-Semantic Supervision. During construction of the multi-prior space, a lightweight codebook decoder provides auxiliary sign-semantic supervision over the vocabulary observed in the

training data. The decoder shapes the representation through vocabulary-level supervision Raffel et al. (2020). A downstream translation head maps the learned SignXray features directly to text through the encoder's continuous feature interface.

3.3 Learning Video-to-Pose Representations

To process raw video inputs directly, we develop a Video2Pose module based on a video ViT Dosovitskiy et al. (2021). Given an input video $V \in \mathbb{R}^{T\times H\times W\times 3}$, the Video2Pose module extracts frame-level features and models their temporal relationships:

$$V_{\text{frame}} = v_1, v_2, ..., v_T = \text{ViT}_{\text{spatial}}(V) \tag{3}$$

where $v_t \in \mathbb{R}^{d_v}$ represents the feature of frame t extracted by the spatial ViT. To capture temporal dependencies, we employ a temporal attention mechanism:

$$V_{\text{temp}} = \text{TemporalAttention}(V_{\text{frame}}) \tag{4}$$

The Video2Pose module projects these temporal features directly into the unified space defined by the frozen Stage-1 pose encoder:

$$\begin{aligned} Z_{\text{pose}} &= E_{\text{pose}}(P_1, \ldots, P_5), \\ \hat{Z} &= \text{Projection}(V_{\text{temp}}). \end{aligned} \tag{5}$$

During Stage 2, Z_{pose} serves as the frozen teacher target and $\hat{Z}$ is the RGB encoder output. The resulting representation preserves the complementary knowledge learned from the five prior sources. During deployment, SignXray produces $\hat{Z}$ directly from RGB through a single encoder Saunders et al. (2021b); Tarrés et al. (2023); Chen et al. (2020); Xie et al. (2021; 2022).

The alignment analysis associates each output feature with a local video window of approximately 5–10 input frames. The resulting feature grid preserves the fused multi-prior information, while any further temporal reduction is performed by the selected downstream head.

3.4 Sign-Aware Latent Refinement

The unified targets provide rich human-centric structure. During downstream CSLR adaptation, two representation-level operations preserve information useful for sign understanding. First, gloss-aware distillation aligns the encoded sequence with sign semantics. A recognition teacher produces logits y^{teacher}. A lightweight projection of the SignXray features produces y^{student}. We optimize

$$\mathcal{L}_{\text{refine}} = \lambda_{\text{KD}}\, \text{KL}\big(y^{\text{student}} \,\|\, y^{\text{teacher}}\big) + \lambda_{\text{CTC}}\, \mathcal{L}_{\text{CTC}}, \tag{6}$$

This downstream objective preserves temporally aligned sign information while inference consumes the continuous SignXray features directly.

Second, variance-based dimension masking reduces redundant channels and preserves the fixed interface expected by downstream heads. For an encoded feature $z_t \in \mathbb{R}^{2048}$, the mask is

$$\hat{z}_t = z_t \odot M, \qquad M_d = \mathbf{1}\Big[\text{Var}(z^{(d)}) \geq \tau\Big]. \tag{7}$$

The active dimensions form a compact, information-dense representation. Masking preserves the tensor shape required for compatibility. Detailed feature compilation, training, and ablations are provided in the appendix.

3.5 Plug-and-Play Encoder Interface

The five prior extractors and the semantic supervision decoder act as training-time teachers. Their knowledge is distilled into and retained by SignXray. At inference time, SignXray exposes a single RGB-to-feature interface,

$$Z = E_{\text{SignXray}}(V), \qquad \hat{Y} = H_{\text{task}}(Z), \tag{8}$$

where H_{task} can be a recognition or translation head. This separation makes the encoder reusable: the heterogeneous priors shape its representation during training, and downstream inference requires RGB video together with the selected task head. The appendix presents our encoder-adapted recognition and translation model as an incremental contribution. The main paper focuses on learning the encoder itself.

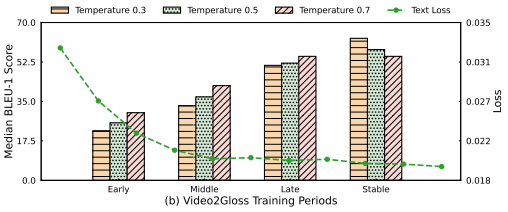


Figure 3: **Latent-space temperature ablation.** Temperature changes the discriminative quality of the representation across training periods; moderate values best balance generation precision and representation stability.

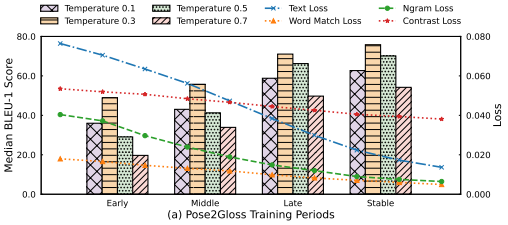
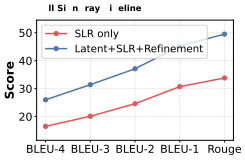
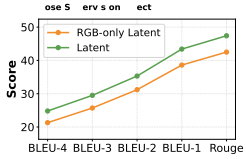


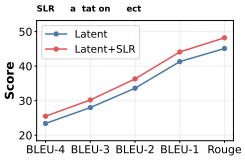
Figure 4: **Optimization of the multi-prior representation.** Text and word-matching losses decline jointly as sign semantics are encoded into the unified space; the text-generation (Ngram) and contrastive objectives stabilize as the heterogeneous features become aligned.



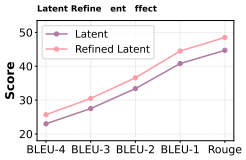
(a) SignXray pipeline added to the RGB-input CSLR baseline



(b) Multi-prior supervision added to the RGB-only latent encoder



(c) Encoder-adapted CSLR head added to SignXray features



(d) Sign-aware refinement applied to the initial SignXray features

Figure 5: **Encoder ablation results.** Performance under a fixed downstream adaptation protocol on the ASLLRP development set. (a) Accuracy contributed by the SignXray representation. (b) Effect of multi-prior supervision. (c) Effect of attaching the encoder-adapted task head. (d) Effect of sign-aware latent refinement. More results are available in Sec. 4.4 and Appendix Sec. D.1.

4 Experimental Results

We evaluate SignXray as a reusable visual encoder across continuous and isolated sign language recognition and sign language translation. The evaluation spans seven benchmarks, multiple downstream architectures, and American, German, Chinese, and Swiss sign languages. Controlled encoder replacements consistently improve the corresponding baselines, while the complete systems achieve state-of-the-art performance on the core benchmarks. These results demonstrate that the unified face-hand-body multi-prior representation transfers effectively across tasks, model families, datasets, and languages. The main paper presents the principal comparisons, encoder ablations, qualitative results, and efficiency analysis; Appendix Sec. E.2 provides expanded cross-architecture and multilingual comparisons in Tables 8–10. Dataset and metric details are reported in Appendix Sec. E.1, and downstream-head details in Appendix Sec. C.

4.1 Ablation Assessment

To further refine the constructed latent space, we study temperature sensitivity and training convergence in Figs. 3 and 4. Across temperatures 0.1, 0.3, 0.5, and 0.7, a moderate value of 0.3 produces the most robust representation during initial space formation. A lower temperature improves precision during later video-to-latent refinement. The convergence curves provide complementary evidence: the text and word-matching losses decline together, and the text-generation (Ngram) and contrastive objectives stabilize as heterogeneous prior features become aligned. Together, these diagnostics show how generation precision, semantic supervision, and cross-prior consistency jointly shape the compact representation.

Figure 5 quantifies the contributions of the SignXray representation and its evaluation head on ASLLRP. Multi-prior-supervised features provide the central gain, and the task adapter and sign-aware refinement provide additional improvements. These results identify the encoder as the primary methodological contribution and the adapted recognition model as a supporting downstream component.

ASLLRP (American Sign Language)
Generated Text: IF STUDENT IX HEAR/LISTEN TEACH+AGENT FUTURE LEARN SOMETHING/ONE *Ground Truth:* IF STUDENT IX HEAR/LISTEN TEACH+AGENT FUTURE LEARN SOMETHING/ONE
Generated Text: BOX/ROOM IX NOT-YET ARRIVE IX SHOULD CONTACT ns-fs-FEDEX *Ground Truth:* BOX/ROOM IX NOT-YET ARRIVE IX SHOULD CONTACT ns-fs-FEDEX
Generated Text: PEOPLE ARRIVE LATE WHO *Ground Truth:* PERSON ARRIVE LATE WHO
Generated Text: HERE ns-fs-MARY FUTURE PAY/BUY CAR *Ground Truth:* MAYBE ns-fs-MARY FUTURE PAY/BUY CAR
Phoenix-14T (German Sign Language)
Generated Text: ES-BEDEUTET HERBST WOLKE UND UND KOENNEN GEWITTER *English:* It means autumn cloud and and can thunderstorm *Ground Truth:* ES-BEDEUTET VIEL WOLKE UND KOENNEN REGEN GEWITTER KOENNEN *English:* That means lots of clouds and again and again strong showers and thunderstorms
Generated Text: TAG DANN SUED SUEDOST NACH MITTAG IX NOVEMBER GEWITTER *English:* Day then south southeast after midday IX November thunderstorm *Ground Truth:* TAG DANN SUED SUEDOST REGEN NACH MITTAG IX AUCH GEWITTER *English:* During the day it rains in the south and south-east, and in the afternoon there are also local thunderstorms
CSL-Daily (Chinese Sign Language)
Generated Text: 这 这 考虑 画画 是 自己 仔细 方法 分析 *English:* This, this, consider painting is one's own careful method analysis *Ground Truth:* 这 画画 是 他 自己 努力 仔细 想 方法 结果 *English:* This painting is the result of his own hard work and careful thinking
Generated Text: 不管 结果 情况 情况 大概 努力 *English:* No matter the result, the situation, the situation, or the effort *Ground Truth:* 不管 结果 情况 我 一直 努力 *English:* No matter the result or situation, I will always work hard

Table 1: **Qualitative evaluation across sign languages.** Examples from ASLLRP, Phoenix-14T, and CSL-Daily show American, German, and Chinese signing, respectively.

4.2 Qualitative Assessment

Table 1 consolidates qualitative results across all three sign languages. In the ASLLRP block, SignXray exactly reproduces the first two sequences, including the long conditional structure and the fingerspelled entity "fs-FEDEX." The third prediction preserves the agentive meaning with "PEOPLE" for "PERSON." The fourth changes the modal "MAYBE" to the spatial term "HERE." These examples show strong structural and lexical preservation, with the remaining ASL errors concentrated in closely related gloss choices and functional modifiers.

On Phoenix-14T, the German predictions preserve central weather concepts such as "WOLKE" (cloud), "GEWITTER" (thunderstorm), "NACH MITTAG" (afternoon), and "SUED SUEDOST." Most errors affect modifiers or nearby lexical choices. On CSL-Daily, the Chinese predictions retain central concepts including "画画" (painting), "仔细" (careful), "不管" (no matter), and "努力" (effort), with omissions and occasional repetition. The unified table provides qualitative evidence across American, German, and Chinese sign-language benchmarks and contrasting domains. We deliberately include challenging cases with common errors in function words, repetition, and fine-grained temporal or spatial modifiers. These localized differences typically preserve the central sign content and sentence-level meaning, showing that SignXray retains stronger qualitative semantic fidelity than n-gram scores alone convey.

Method	Modality		Dev						Test					
	Pose	RGB	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEURT	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEURT
Gloss-free sign language translation														
GloFE-VN Lin et al. (2023)	✓		21.06	12.34	8.68	6.68	21.37	36.75	21.56	12.74	9.05	7.06	21.75	36.35
Conv-GRU† Cihan Camgoz et al. (2018)		✓	16.72	8.95	6.31	4.82	16.25	25.36	16.11	8.85	6.18	4.58	16.10	25.65
I3D-Transformer Shi et al. (2022)		✓	18.26	10.26	7.17	5.60	18.88	29.17	18.31	10.15	7.19	5.66	18.64	28.82
OpenASL Shi et al. (2022)		✓	20.10	11.81	8.43	6.57	20.43	31.22	20.92	12.08	8.59	6.72	21.02	31.09
C^2RL Chen et al. (2024a)		✓	–	–	–	–	–	–	31.46	21.85	16.58	13.21	31.36	–
VSLT (gloss-free) Zhao et al. (2026)		✓	–	–	–	–	–	–	38.60	25.59	19.49	14.35	36.68	–
SignLlama (gloss-free) Gan et al. (2026a)		✓	35.11	26.44	19.02	16.59	35.28	–	37.37	25.97	17.70	15.16	37.74	–
Uni-Sign Li et al. (2025)	✓		49.52	36.06	28.06	22.39	42.64	60.47	49.10	35.91	28.12	22.67	42.77	60.08
Uni-Sign Li et al. (2025)	✓	✓	50.84	37.82	29.83	24.16	44.58	61.28	49.35	36.32	28.55	23.14	43.22	60.40
Uni-Sign + SignXray	✓	✓	**55.06**	**41.16**	**33.13**	**25.85**	**46.60**	**66.41**	**50.75**	**37.60**	**29.72**	**24.92**	**44.75**	**66.62**

Table 2: **Gloss-free SLT results on OpenASL.** Published results are reported from the original studies; Uni-Sign baselines are reproduced from Uni-Sign Li et al. (2025). For Uni-Sign, modality marks follow its original training convention; the SignXray encoder receives only RGB video at inference. † denotes a method reproduced by Shi et al. (2022).

	PHOENIX2014-T										
	Method	Dev					Test				
		Rouge	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	BLEU1	BLEU2	BLEU3	BLEU4
Sign2Text	STMC-T Zhou et al. (2021c)	48.24	47.60	36.43	29.18	24.09	46.65	46.98	36.09	28.70	23.65
	SignBT Zhou et al. (2021b)	50.29	51.11	37.90	29.80	24.45	49.54	50.80	37.75	29.72	24.32
	MMTLB Chen et al. (2022a)	53.10	53.95	41.12	33.14	27.61	52.65	53.97	41.75	33.84	28.39
	VSLT (gloss-free) Zhao et al. (2026)	–	–	–	–	–	49.94	51.33	39.18	31.79	26.83
	SignLlama (gloss-free) Gan et al. (2026a)	52.53	52.13	38.88	30.61	25.73	52.81	52.54	40.96	32.87	26.74
	SLTUNET Zhang et al. (2023a)	52.23	-	-	-	27.87	52.11	52.92	41.76	33.99	28.47
	SLTUNET + SignXray	**55.32**	**55.24**	**43.10**	**35.45**	**30.08**	**55.15**	**55.08**	**42.92**	**35.28**	**29.91**
	TwoStream-SLT Chen et al. (2022b)	54.08	54.32	41.99	34.15	28.66	53.48	54.90	42.43	34.46	28.95
	TwoStream-SLT + SignXray	**54.68**	**54.84**	**42.55**	**34.76**	**29.03**	**53.80**	**55.31**	**42.75**	**34.92**	**29.54**
	CorrNet+ Hu et al. (2024)	54.54	54.56	42.31	34.48	29.13	53.76	55.32	42.74	34.86	29.42
	CorrNet+ + SignXray	**55.17**	**55.03**	**42.83**	**34.81**	**29.66**	**54.29**	**55.81**	**43.26**	**35.44**	**29.76**
	CSL-Daily										
	Method	Dev					Test				
		Rouge	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	BLEU1	BLEU2	BLEU3	BLEU4
Sign2Text	MMTLB Chen et al. (2022a)	53.38	53.81	40.84	31.29	24.42	53.25	53.31	40.41	30.87	23.92
	VSLT (gloss-free) Zhao et al. (2026)	–	–	–	–	–	48.07	48.98	36.34	27.87	21.97
	SignLlama (gloss-free) Gan et al. (2026a)	51.62	52.12	41.23	33.23	24.83	51.43	52.38	39.78	33.91	24.76
	SLTUNET Zhang et al. (2023a)	53.58	-	-	-	23.99	54.08	54.98	41.44	31.84	25.01
	SLTUNET + SignXray	**57.25**	**57.10**	**44.50**	**35.80**	**28.75**	**57.12**	**56.95**	**44.32**	**35.62**	**28.58**
	TwoStream-SLT Chen et al. (2022b)	55.10	55.21	42.31	32.71	25.76	55.72	55.44	42.59	32.87	25.79
	TwoStream-SLT + SignXray	**55.54**	**55.78**	**42.74**	**33.36**	**26.14**	**56.16**	**55.81**	**42.91**	**33.32**	**26.43**
	CorrNet+ Hu et al. (2024)	55.52	55.64	42.78	33.13	26.14	55.84	55.82	42.96	33.26	26.14
	CorrNet+ + SignXray	**56.08**	**56.08**	**43.17**	**33.78**	**26.79**	**56.44**	**56.40**	**43.32**	**33.78**	**26.79**
	Uni-Sign Li et al. (2025)	56.03	55.30	-	-	26.25	56.51	55.08	-	-	26.36
	Uni-Sign + SignXray	**56.40**	**55.78**	**42.53**	**33.51**	**26.58**	**56.83**	**55.74**	**43.95**	**33.86**	**26.82**

Table 3: **SLT comparison** on PHOENIX2014-T and CSL-Daily. Gloss-free rows are published RGB-only results; SignXray paired rows replace the visual encoder while retaining the downstream protocol. −: unavailable.

Continuous SLR (CSLR)			**Isolated SLR (ISLR)**	
Method	RWTH-14T (WER ↓)	CSL-Daily (WER ↓)	Method	WLASL-2000 (P-I ↑)
GEU Tang et al. (2022)	49.9	-	BEST Zhao et al. (2023)	46.25
Bert Guo et al. (2024)	21.2	30.8	SignBERT+ Hu et al. (2023a)	48.85
Bert fine-tuning Guo et al. (2024)	20.4	30.0	MSLU Zhou et al. (2025)	56.29
CorrNet Hu et al. (2023b)	20.5	30.1	NLA-SLR Zuo et al. (2023)	61.05
GPGN Guo et al. (2024)	20.5	30.0	Uni-Sign Li et al. (2025)	63.52
CorrNet+ Hu et al. (2024)	19.1	28.2	Sigma Pu et al. (2025)	64.40
SLTUNET Zhang et al. (2023a)	20.4	27.6	SLTUNET Zhang et al. (2023a)	64.82
SLTUNET + SignXray	**18.6**	**24.3**	**SLTUNET + SignXray**	**68.29**

Table 4: **Continuous and isolated SLR results.** WER is reported for RWTH-2014-T and CSL-Daily, and P-I accuracy for WLASL-2000. The final rows compare SLTUNET with SignXray.

4.3 Comparison with Previous Work and Plug-and-Play Transfer

Table 2 reports OpenASL transfer, Table 3 evaluates paired SLT configurations on PHOENIX2014-T and CSL-Daily, and Table 4 covers continuous and isolated recognition with BLEU, ROUGE, BLEURT, WER, and P-I accuracy.

On OpenASL, adding SignXray to pose-and-RGB Uni-Sign improves every development and test metric, including BLEU-4 from 24.16 to 25.85 on the development set and from 23.14 to 24.92 on the test set. Table 3 demonstrates similarly consistent plug-and-play gains across three SLT architectures on PHOENIX2014-T and four configurations on CSL-Daily. It also places these paired results alongside recent RGB-only gloss-free systems, VSLT and SignLlama, on both benchmarks. Table 4 further records state-of-the-art continuous and isolated recognition performance, confirming that the representation supports both sequence prediction and video-level classification.

Appendix Sec. E.2 expands this evaluation with three comparisons. Table 8 isolates the encoder replacement across C2SLR, CorrNet+, and SLTUNET on ASLLRP; Table 9 evaluates SignXray with Uni-Sign on How2Sign; and Table 10 tests ASL-only and multilingual YouTube-SL-25 configurations across How2Sign and three SignSuisse languages. Together with the state-of-the-art results in tables 2 to 4, the consistent gains across architectures, datasets, languages, and metrics validate SignXray as a powerful plug-and-play encoder for diverse sign language understanding systems.

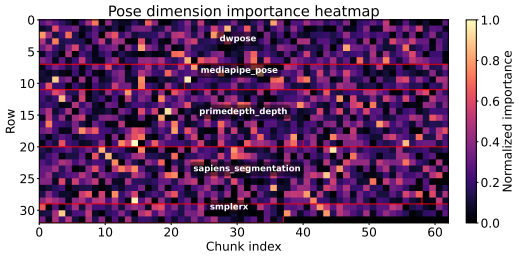


Figure 6: **Dimension importance across the five priors.** Normalized importance over 1,959 input dimensions, with red lines separating SMPLer-X, DWPose, MediaPipe, PrimeDepth, and Sapiens. Warmer colors indicate greater activation importance.

Removed Prior Source	WER (%)	Δ WER	Relative Degradation
SignXray (Full Modality)	26.59	–	–
w/o SMPLer-X	29.65	+3.06	11.51%
w/o DWPose	32.31	+5.72	21.51%
w/o Mediapipe	29.37	+2.78	10.46%
w/o PrimeDepth	30.89	+4.30	16.17%
w/o Sapiens	30.53	+3.94	14.82%

Table 5: **Contribution of each prior source.** Each row is trained with the indicated prior source removed and optimized to convergence under the same recipe. Ten random seeds are used for each ablation, with the best development checkpoint selected for test evaluation. The degradation after removal highlights the complementary information consolidated by SignXray.

4.4 Contribution of the Five Prior Sources

Figure 6 visualizes activation importance across the 1,959 input dimensions. Important dimensions span all five source regions, indicating that the fused encoder uses heterogeneous information throughout the multi-prior input. Table 5 provides a controlled source-level ablation. We retrain the Vid2Pose variant from initialization with one prior source removed, using the same CSLR recipe through convergence. Every source removal increases WER: the MediaPipe ablation adds 2.78 points, and the DWPose ablation adds 5.72 points. PrimeDepth and Sapiens also yield substantial gains beyond the three pose estimators. Together, these results support the encoder's central design: the five sources contribute complementary information that is jointly consolidated in a single representation. These degradations and residual gains indicate both the value and partial redundancy of the priors, enabling SignXray to retain sign-relevant cues in a compact reusable space.

4.5 Inference Efficiency Assessment

Method	FPS ↑	Power (W) ↓	WER (%) ↓
SLTUNET Baseline (RGB Input)	0.05	26.22	43.26
SignXray (SLTUNET in Latent Space)	2.42	24.58	26.59
Uni-Sign Baseline (RGB Input)	3.76	44.07	40.87
SignXray (Uni-Sign in Latent Space)	19.27	38.11	23.66

Table 6: **Inference efficiency on ASLLRP.** SignXray reaches 2.42 FPS in latent space and obtains the lowest WER among the evaluated variants.

Table 6 compares two baseline–SignXray pairs on ASLLRP. Each baseline operates on raw RGB visual features, whereas its SignXray counterpart uses compact features from our multi-prior-trained encoder. For SLTUNET, SignXray reaches `2.42 FPS`, corresponding to a 48.4× acceleration over the 0.05 FPS baseline. For Uni-Sign, it reaches `19.27 FPS`, or approximately 5.1× the 3.76 FPS baseline. The Uni-Sign pairing also attains the lowest WER of `23.66%`. These results demonstrate the efficiency and recognition quality of the compact multi-prior representation across both downstream configurations.

5 Conclusion

This paper presents **SignXray**, a plug-and-play sign video encoder learned in a unified Face-Hand-Body multi-prior space. During training, SignXray consolidates complementary body, hand, face, depth, and segmentation supervision and distills it into a single RGB ViT Dosovitskiy et al. (2021). At inference time, the RGB encoder directly produces the information-dense representation through a common feature interface, with no external prior extractor required.

Across continuous recognition, isolated recognition, and translation benchmarks, encoder-adapted heads demonstrate the representation's utility for multiple forms of sign understanding and lower inference cost relative to the evaluated pixel-space baseline. These heads serve as evaluation adapters, and their complete architecture is retained in the appendix as an incremental contribution. The separation between multi-prior training and RGB-only inference supports direct integration of SignXray into additional downstream models and larger-scale sign-language settings.

AI USE STATEMENT

In this work, generative AI tools were used to polish the manuscript prose and assist with preparing explanatory illustrations. Generative AI tools were not used to design the model. All AI-assisted text and visual materials were reviewed and edited by the authors. The authors take responsibility for the final content of this paper, including all text, claims, and artifacts produced with AI assistance.

REFERENCES

Sobhan Asasi, Ozge Mercanoglu Sincan, and Richard Bowden. Signet: Motion-level knowledge transfer for cross-language sign language translation, 2026. URL https://arxiv.org/abs/2606.28626.

Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 182–191, January 2022.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation, 2023.

Necati Cihan Camgöz, 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 (CVPR)*, 2018.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody Dance Now. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2019.

James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. Automatic and Efficient Human Pose Estimation for Sign Language Videos. *International Journal of Computer Vision (IJCV)*, 2014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5120–5130, 2022a.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *NeurIPS*, 2022b.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. C^2rl: Content and context representation learning for gloss-free sign language translation and retrieval. *arXiv*, 2024a.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. Factorized learning assisted with large language model for gloss-free sign language translation. In *LREC-COLING*, pp. 7071–7081, 2024b.

Hung-Yueh Chiang, Yen-Liang Lin, Yueh-Cheng Liu, and Winston H Hsu. A unified point-based framework for 3d segmentation. In *3DV*, 2019.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pp. 1931–1942. PMLR, 2021.

Anal Roy Chowdhury and Debarshi Kumar Sanyal. Can small vision–language models perform sign language translation? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32150–32166, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1609. URL https://aclanthology.org/2026.findings-acl.1609/.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pp. 7784–7793, 2018.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Helen Cooper and Richard Bowden. Large Lexicon Detection of Sign Language. In *International Workshop on Human-Computer Interaction*, 2007.

Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

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 (CVPR)*, 2021.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. Signllm: Sign language production large language models, 2025a. URL https://arxiv.org/abs/2405.10718.

Sen Fang, Yalin Feng, Hongbin Zhong, Yanxin Zhang, and Dimitris N. Metaxas. Stable signer: Hierarchical sign language generative model, 2025b. URL https://arxiv.org/abs/2512.04048.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. Signdiff: Diffusion model for american sign language production, 2025c. URL https://arxiv.org/abs/2308.16082.

Sen Fang, Hongbin Zhong, Yanxin Zhang, and Dimitris N. Metaxas. Signverse-2m: A two-million-clip pose-native universe of 55+ sign languages, 2026. URL https://arxiv.org/abs/2605.01720.

Liqian Feng, Lintao Wang, Xiaochen Liu, Anusha Withana, Ken-Tye Yong, Dehui Kong, Zhiyong Wang, and Kun Hu. Q-bridgenet: A quantization network for cross-lingual sign language translation, 2026. URL https://arxiv.org/abs/2607.11215.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2012.

Shiwei Gan, Xiao Liu, Yafeng Yin, Zhiwei Jiang, Bowen Guo, Lie Xie, Sanglu Lu, and Hongkai Wen. Signllama: Enhancing gloss-free sign language translation by prioritizing visual features for llms, 2026a. URL https://arxiv.org/abs/2608.09006.

Shiwei Gan, Xiao Liu, Yafeng Yin, Nan Liu, Kuizhuang Liu, Desibieer Tuerdaken, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Learning effective sign features without text for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9827–9836, June 2026b.

Kirsti Grobel and Marcell Assan. Isolated Sign Language Recognition using Hidden Markov Models. In *IEEE International Conference on Systems, Man, and Cybernetics*, 1997.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22506–22521, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1157. URL https://aclanthology.org/2025.findings-acl.1157/.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28792–28810, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1397. URL https://aclanthology.org/2025.acl-long.1397/.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306, 2018.

Leming Guo, Wanli Xue, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Dimitris Metaxas. Gloss prior guided visual feature learning for continuous sign language recognition. *IEEE Transactions on Image Processing*, 33:3486–3495, 2024. doi: 10.1109/TIP.2024.3404869.

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 (CVPR)*, pp. 28828–28837, June 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239, 2023a.

Junyi Hu, Zhewen He, Haomian Huang, Aoxiang Yang, and Yi Fang. Vtamo: Video-text alignment model for sign language translation, 2026. URL https://arxiv.org/abs/2607.09126.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023b.

Lianyu Hu, Wei Feng, Liqing Gao, Zekang Liu, and Liang Wan. Cornet+: Sign language recognition and translation via spatial-temporal correlation, 2024. URL https://arxiv.org/abs/2404.11111.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards Fast and High-Quality Sign Language Production. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

Yiyang Jiang, Li Zhang, Xiao-Yong Wei, and Li Qing. Think in latent thoughts: A new paradigm for gloss-free sign language translation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9993–10012, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.454. URL https://aclanthology.org/2026.acl-long.454/.

Timor Kadir, Richard Bowden, Eng-Jon Ong, and Andrew Zisserman. Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2004.

Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust. Speech Signal Process.*, 35(3):400–401, 1987.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024.

Youngmin Kim, Minji Kwak, Dain Lee, Yeongeun Kim, and Hyeongboo Baek. Keypoint based sign language translation without glosses. *arXiv preprint arXiv:2204.10511*, 2022.

Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1):19–28, 2002. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(01)00041-3. URL https://www.sciencedirect.com/science/article/pii/S0167639301000413.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683, 2019a.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural Sign Language Translation based on Human Keypoint Estimation. *Applied Sciences*, 2019b.

Oscar Koller. Quantitative Survey of the State of the Art in Sign Language Recognition. *arXiv preprint arXiv:2008.09918*, 2020.

Oscar Koller, Jens Forster, and Hermann Ney. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding (CVIU)*, 2015.

Chang Hao Lai, Rui Zhao, Xuewen Zhong, Jinsong Su, and Yidong Chen. Selective contrastive learning for gloss free sign language translation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45648–45660, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2116. URL https://aclanthology.org/2026.acl-long.2116/.

Dongxu Li, Cristian Rodriguez, 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, 2020a.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 1459–1469, 2020b.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*, 2025.

Scott K Liddell et al. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. GLoFE: gloss-free end-to-end sign language translation. In *ACL*, pp. 12904–12916, 2023.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM TOG*, 2015.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A Unified Model for Vision, Language, and Multi-modal Tasks. February 2023a. URL https://openreview.net/forum?id=E01k9048soZ.

Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. Multimodal procedural planning via dual text-image prompting. *CoRR*, abs/2305.01795, 2023b.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pp. 23803–23828. PMLR, 2023.

D. Metaxas. Deformable model and hmm-based tracking, analysis and recognition of gestures and faces. In *Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No.PR00378)*, pp. 136–140, 1999. doi: 10.1109/RATFG.1999.799236.

Junwen Mo, MinhDuc Vo, Noriki Nishida, Shin'ichi Satoh, and Hideki Nakayama. J-shuwa: A large-scale web-collected Japanese Sign Language-Japanese parallel corpus. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36559–36574, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1821. URL https://aclanthology.org/2026.findings-acl.1821/.

Amit Moryossef. Complete toolkit for working with poses. https://github.com/AmitMY/pose-format/, 2022.

Amit Moryossef, Ioannis Tsochantaridis, Roee Yosef Aharoni, Sarah Ebling, and Srini Narayanan. Real-time sign language detection using human pose estimation. 2020. https://www.slrtp.com/.

Carol Neidle and Carey Ballard. Why alternative gloss labels will increase the value of the wlasl dataset. 2022.

Carol Neidle and Dimitris Metaxas. American sign language linguistic research project (asllrp) sign bank. https://dai.cs.rutgers.edu/dai/s/signbank, June 2025. ASLLRP Continuous Signing Corpora, version 3.

Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. NEW Shared & Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank. In *Language Resources and Evaluation. 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. Miyazaki, Japan, 2018-05-12*, pp. 147–154.

Carol Neidle, Augustine Opoku, and Dimitris Metaxas. ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP), 2022a.

Carol Neidle, Augustine Opoku, and Dimitris Metaxas. Asl video corpora & sign bank: Resources available through the american sign language linguistic research project (asllrp), 2022b. URL https://arxiv.org/abs/2201.07899.

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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Muxin Pu, Mei Kuan Lim, Chun Yong Chong, and Chen Change Loy. Sigma: Semantically informative pre-training for skeleton-based sign language understanding, 2025. URL https://arxiv.org/abs/2509.21223.

Shi Qiu, Saeed Anwar, and Nick Barnes. Dense-resolution network for point cloud classification and segmentation. In *WACV*, pp. 3813–3822, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. Towards privacy-aware sign language translation at scale. In *ACL*, pp. 8624–8641, 2024.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020a.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020b.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision (IJCV)*, 2021a.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021b.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Skeletal Graph Self-Attention: Embedding a Skeleton Inductive Bias into Sign Language Production. *arXiv preprint arXiv:2112.05277*, 2021c.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. In *EMNLP*, pp. 6365–6379, 2022.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*, 2020.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014a.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014b.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia*, 24:4433–4445, 2022. doi: 10.1109/TMM.2021.3117124.

Garrett Tanzer and Biao Zhang. Youtube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=nFVsK3QLgs.

Laia Tarrés, Gerard I. Gàllego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. Sign language translation from instructional videos. In *CVPRW*, pp. 5625–5635, 2023.

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023.

Dave Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. In *NeurIPS*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Špela Vintar, Boštjan Jerko, and Marjetka Kulovec. Compiling the slovene sign language corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference (LREC)*, volume 5, pp. 159–162, 2012.

C. Vogler and D. Metaxas. Asl recognition based on a coupling between hmms and 3d motion analysis. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 363–369, 1998. doi: 10.1109/ICCV.1998.710744.

U. Von Agris and K.-F. Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *Workshop on Representation and Processing of Sign Languages*, pp. 243–246, 2010.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2018a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Signrep: Enhancing self-supervised sign representations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22804–22814, 2025. doi: 10.1109/ICCV51701.2025.02117.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

Lujia Yang, Weicai Yan, Yongbo He, Qifei Zhang, Tao Jin, Jinshan Zhang, Meng Xi, and Jianwei Yin. SAME: Signer-aware mixture-of-experts for test-time adaptation in sign language translation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings*

of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 21274–21289, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.973. URL https://aclanthology.org/2026.acl-long.973/.

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*, pp. 4210–4220, 2023.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5109–5119, 2022.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014.

Denis Zavadski, Damjan Kalšan, and Carsten Rother. Primedepth: Efficient monocular depth estimation with a stable diffusion preimage, 2024. URL https://arxiv.org/abs/2409.09144.

Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 3395–3403, 2020a.

Jan Zelinka and Jakub Kanis. Neural Sign Language Synthesis: Words Are Our Glosses. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020b.

Biao Zhang, Mathias Müller, and Rico Sennrich. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=EBS4C77p_5S.

Feihu Zhang, Jin Fang, Benjamin Wah, and Philip Torr. Deep fusionnet for point cloud semantic segmentation. In *ECCV*, 2020.

Sen Zhang, Xiaoxiao He, Di Liu, Zhaoyang Xia, Mingyu Zhao, Chaowei Tan, Vivian Li, Bo Liu, Dimitris N. Metaxas, and Mubbasir Kapadia. Large sign language models: Toward 3d american sign language translation, 2025. URL https://arxiv.org/abs/2511.08535.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923, 2023b.

Rui Zhao, Liang Zhang, Biao Fu, Ruiquan Zhang, Yidong Chen, and Xiaodong Shi. Variational sign language translation. *International Journal of Computer Vision*, 134(9):408, 2026. doi: 10.1007/s11263-026-02978-x.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. Best: Bert pre-training for sign language recognition with coupling tokenization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3597–3605, 2023.

Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1316–1325, 2021a.

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, 2021b.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779, 2021c.

Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. Scaling up multimodal pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yang Zhou, Zhaoyang Xia, Yuxiao Chen, Carol Neidle, and Dimitris N. Metaxas. A multimodal spatio-temporal GCN model with enhancements for isolated sign recognition. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Johanna Mesch, and Marc Schulder (eds.), *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pp. 408–419, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.signlang-1.45/.

Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*, pp. 5131–5140, 2022.

Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14890–14900, 2023.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23806–23816, October 2025.

A Background Information

A.1 Diffusion Models

Diffusion models have achieved breakthroughs in the generative arena in recent years. From the initial DDPM Ho et al. (2020) to Stable Diffusion Rombach et al. (2021), diffusion models have demonstrated powerful image generation capabilities Saunders et al. (2021c); Wang et al. (2018b). Researchers have also begun applying diffusion to discrete text sequence generation, with promising results Ko et al. (2019b); Boháček & Hrúz (2022). Particularly in cross-modal conversion tasks, diffusion models have shown excellent performance, providing important references for converting SL videos to text Saunders et al. (2020b); Cooper & Bowden (2007); Zhou et al. (2024).

Latent diffusion models perform generation in a compact latent space Rombach et al. (2022), separating high-dimensional visual synthesis from lower-dimensional latent modeling. This perspective has also informed sign-language generation and representation learning Fang et al. (2025b;c; 2026); Gueuwou et al. (2025b); Li et al. (2025). SignXray extends it to sign understanding: its information-dense features shorten the sequence presented to downstream recognition and translation heads, while its RGB encoder compiles complementary multi-prior knowledge into every feature.

A.2 Recent Sign Language Research

Gloss-based representations provide an established interface for modeling sign sequences and their linguistic organization Camgöz et al. (2018); Cihan Camgöz et al. (2020); Liddell et al. (2003); Zelinka & Kanis (2020a;b); Neidle & Ballard (2022); Zhou et al. (2024); Neidle et al. (2022b); Zhang et al. (2025).

Complementary recent work studies signer-aware test-time adaptation, selective contrastive alignment, visual sign features learned without text, and compact vision–language models Yang et al. (2026); Lai et al. (2026); Gan et al. (2026b); Chowdhury & Sanyal (2026). J-Shuwa broadens publicly available Japanese sign-language resources, while Signs as Tokens develops retrieval-enhanced multilingual sign generation Mo et al. (2026); Zuo et al. (2025). These systems develop specialized alignment, adaptation, or translation heads; SignXray supplies a reusable visual interface that can strengthen such downstream designs. Its heterogeneous 2D and 3D pose, depth, and segmentation priors are distilled into a single RGB encoder without running the prior extractors during inference. Appendix Sec. E reports complementary cross-architecture, multilingual, and concurrent-system benchmarks under a controlled paired encoder-substitution protocol.

B Methodology for Building Sign Language Latent Space

Dataset Preparation. Our process of constructing the rich pose latent space utilize a linguistically annotated ASL dataset of continuous signing: the ASLLRP SignStream® 3 Corpus Neidle et al. (2022a). This dataset contains over 80 hours of American Sign Language (ASL) videos with synchronized front view, side view, and facial close-up recordings. These videos have been linguistically annotated using SignStream® Neidle et al. software, providing detailed manually annotated gloss labels, sign types, handshape information, and non-manual grammatical markers. We use the official ASLLRP split of 1,686 training, 210 development, and 212 test utterances.

For video processing, we apply five pose extraction models: SMPLer-X Cai et al. (2023) for accurate 3D human body parameters, DWPose Yang et al. (2023) for real-time 2D keypoint detection, MediaPipe Lugaresi et al. (2019) for lightweight 3D pose prediction, PrimeDepth Zavadski et al. (2024) for hand depth information, and Sapiens segmentation Khirodkar et al. (2024) for fine-grained body part segmentation. Our preprocessing retains incomplete and partially occluded samples to expose the encoder to real-world quality variation. Videos are standardized to 30fps and processed at uniform spatial resolution. For each official split, we extract multimodal pose features with the same fixed preprocessing. Only training-split features enter teacher optimization and the estimation of vocabulary, normalization statistics, and masking parameters.

Sign recognition presents significant challenges in multimodal representation learning, particularly in converting continuous visual gestures into discrete textual descriptions Cihan Camgöz et al. (2020); Cihan Camgoz et al. (2018); Saunders et al. (2020a); Kim et al. (2022). Regarding the

construction of the posture latent space, We propose a 2-stage training approach based on ViT that systematically learns Pose2Gloss and Video2Pose stages.

B.1 Stage 1: ViT Pose2Gloss Part Training

CodeBook Decoder. The *Custom CodeBook* in Fig. 2 is a trainable word-index vocabulary that provides auxiliary semantic supervision for Stage 1. For each training configuration, we construct $\mathcal{V}_{\text{train}}$ from annotated training-split tokens and add <pad>, <eos>, and <unk>; the resulting fixed table defines both token embeddings and the output vocabulary of the Stage-1 decoder. A learned decoder maps the $T \times 2{,}048$ Stage-1 features to logits over $|\mathcal{V}_{\text{train}}|$ tokens (with maximum length 64), and the same table supplies the targets for the word-matching and contrastive objectives below. The table remains fixed within a run and is constructed independently from each configuration training partition. It shapes the Stage-1 teacher, while Stage 2 and downstream task heads consume the shared 2,048-dimensional feature interface.

Text Generation Loss. Following seq2seq principles Sutskever et al. (2014b), we employ a cross-entropy text generation loss Mao et al. (2023) that measures the probabilistic divergence between predicted and ground truth token (word index in codebook) sequences:

$$L_{\text{text}} = -\sum_{i=1}^{N} y_{\text{true},i} \log(y_{\text{pred},i}) \tag{9}$$

where N denotes sequence length, $y_{\text{true},i}$ represents ground truth tokens, and $y_{\text{pred},i}$ represents predicted token probabilities. The Ngram Loss label in Fig. 4 denotes this sequence-level text-generation objective.

Word Matching Loss. To further improve vocabulary precision and text coherence, we introduce a word matching loss $\mathcal{L}_{\text{word}}$ that directly compares predicted and ground-truth token embeddings using cosine similarity. This loss is defined as:

$$\mathcal{L}_{\text{word}} = \frac{1}{B} \sum_{b=1}^{B} (1 - \text{sim}(e_{\text{pred},b}, e_{\text{true},b})), \tag{10}$$

where B is the batch size, $e_{\text{pred},b}$ and $e_{\text{true},b}$ represent the embeddings of the predicted and ground-truth tokens for batch sample b, respectively, and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function. By minimizing this loss, the model learns to align the predicted token embeddings with their true counterparts, enhancing the semantic accuracy and fluency of the generated text. This approach proves particularly effective in addressing over-long or incoherent outputs observed with the text generation loss alone.

Contrastive Learning Loss. Inspired by recent multimodal representation learning Chen et al. (2020); Lu et al. (2023b); Zhang et al. (2023b), we introduce a contrastive alignment loss that encourages semantic proximity between pose and text embeddings:

$$L_{\text{contrast}} = -\log \frac{\exp(sim(z_{\text{pose}}, e_{\text{text}})/\tau)}{\sum_j \exp(sim(z_{\text{pose}}, e^j_{\text{text}})/\tau)}. \tag{11}$$

The loss leverages cosine similarity between pose embeddings z_{pose} (output of module) and text embeddings e_{text} (index of codebook), with τ as a temperature scaling parameter to control embedding separation, where $\exp(\cdot)$ denotes the exponential function.

Composite Loss Formulation. The total loss is formulated as a weighted combination of these components:

$$\begin{aligned} L_{\text{total}} = & \lambda_{\text{text}} L_{\text{text}} \\ & + \lambda_{\text{word}} L_{\text{word}} + \lambda_{\text{contrast}} L_{\text{contrast}}. \end{aligned} \tag{12}$$

Config	Stage 1 (Pose2Gloss)	Stage 2 (Video2Pose)
optimizer	AdamW	AdamW
base learning rate	1×10^{-3}	1×10^{-3}
weight decay	0.01	0.01
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$
learning rate schedule	Cosine decay (min = $0.05 \times lr$)	Cosine decay (min = $0.01 \times lr$)
warmup steps	Dynamic (2%-10% of total steps)	10% of total steps
training epochs	500–1000	300
batch size	32	32
gradient accumulation	4	1
gradient clipping	7.0	1.0
label smoothing	0.1	N/A
teacher forcing ratio	Decay from 0.5 to 0	N/A
text loss weight	1	N/A
word match weight	1	N/A
contrast loss weight	1	N/A
max sequence length	64	N/A
temperature	0.7 (inference)	N/A
beam size	5 (inference)	N/A
length penalty	0.7 (inference)	N/A
top-k	50 (sampling)	N/A
top-p	0.92 (sampling)	N/A
repetition penalty	2.5 (beam search)	N/A

Table 7: **Comprehensive training configurations and hyperparameters:** Optimizer settings, learning schedules, training parameters, and inference configurations for both stages.

B.2 Stage 2: ViT Video2Pose Part Training

Following feature transfer learning principles Yosinski et al. (2014), we freeze the Stage-1 pose encoder and train Video2Pose to reproduce its unified representation:

$$L_{\text{video2pose}} = \|\hat{Z} - Z_{\text{pose}}\|_2^2, \tag{13}$$

where $Z_{\text{pose}} = E_{\text{pose}}(P_1, \ldots, P_5)$ is the frozen multi-prior teacher target and $\hat{Z} = E_{\text{video}}(V)$ is the representation predicted from RGB video. This 2-stage approach transfers the unified multi-prior space into the RGB encoder used by downstream models. The RGB sequence is projected onto the teacher feature grid before the feature-level objective is evaluated. The prior extractors and semantic decoder are therefore training-time teachers; their one-time preprocessing and training cost is separate from deployment, which uses the RGB encoder and a selected downstream head without running those extractors. The source-specific encoders concatenate to $F \in \mathbb{R}^{T \times 1959}$, and PadMatch maps this sequence to the fixed $T \times 2048$ interface before the ViT layers. The same temporal grid is used for $\hat{Z}$ after RGB projection; downstream heads may apply their own feature projection, such as the 512-dimensional recurrent state shown in Fig. 7. The active-channel mask is estimated from training features and is applied elementwise without changing the external 2048-dimensional shape. Further temporal-compression details are analyzed in Appendix Sec. D.2.

B.3 Implementation Details of Latent Space

For Stages 1 and 2, we use the linguistically annotated ASLLRP SignStream® 3 Corpus of continuous signing Neidle et al. (2022a) as primary datasets for American Sign Language (ASL) recognition. During Stage 1, we initially found that the heterogeneity and varying formats of pose features from diverse sources had a significant negative impact on the structural integrity of our latent space. To address this, we prioritized optimizing the cross-modal alignment and feature consistency in the Pose2Gloss module (Stage 1) through a gloss-aware latent distillation loss. Feature-level distillation alone left substantial variation in sign-to-text generation quality.

We explored text generation using a cross-entropy loss $\mathcal{L}_{\text{text}}$. This objective improved basic text output and revealed over-long sequences that motivated additional vocabulary supervision. Direct vocabulary alignment produced the strongest preliminary results. We therefore introduced a word

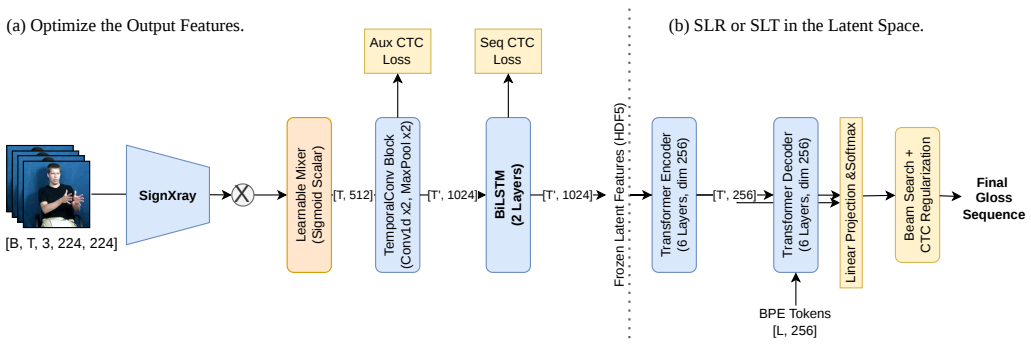


Figure 7: **Encoder-adapted downstream sign understanding.** SignXray features are organized and passed to task-specific temporal heads for continuous recognition or translation. The downstream model is an incremental contribution. The main paper focuses on learning the reusable multi-prior encoder.

matching loss that measures cosine similarity between predicted and ground-truth token embeddings, encouraging precise vocabulary generation within our custom CodeBook. Additionally, we incorporated contrastive learning with temperature parameter τ to further encourage semantic proximity between the ViT-constructed pose latent space and text representations.

For Stage 2 (Video2Pose), we freeze the pose encoder trained in Stage 1 and optimize the RGB ViT with $\mathcal{L}_{\text{video2pose}} = \|\hat{Z} - Z_{\text{pose}}\|_2^2$. This objective transfers the unified multi-prior representation into a direct RGB-to-feature mapping. Stage 2 uses this feature-regression objective without a text decoder or downstream KD/CTC loss. These optimizations, combined with gradient accumulation, as detailed in Table 7, resolved initial training instability and enhanced model performance. The best options for weights, pose types, and temperature values for different losses were determined through extensive ablation experiments.

C Encoder-Adapted Sign Recognition and Translation

The SignXray encoder is the primary contribution of this paper. We additionally develop task heads that demonstrate complete sign-understanding systems built on its output. This downstream architecture constitutes an incremental contribution, with its recognition, translation, optimization, and decoding details provided in the appendix. Both tasks consume SignXray features through an RGB-only inference pipeline.

C.1 Task-Specific Adaptation

Let $Z = E_{\text{SignXray}}(V)$ denote the sequence produced by the Stage-2 RGB encoder; the Stage-1 multi-prior teacher remains frozen. For continuous recognition, a temporal head maps Z to gloss probabilities and is optimized with CTC and sequence-level supervision. For translation, the same encoder features are consumed directly by a sequence-to-sequence text decoder:

$$\hat{Y}_{\text{text}} = D_{\text{SLT}}(Z), \qquad \mathcal{L}_{\text{SLT}} = -\log p(Y_{\text{text}} \mid Z). \tag{14}$$

The translation path maps continuous SignXray features directly to text. The CSLR KD and CTC objectives belong to the downstream adaptation stage; Stage 2 RGB training uses only the feature-regression objective defined in Appendix Sec. C. The following subsections provide the complete CSLR instantiation together with the shared feature organization and training details; the translation head reuses the same SignXray interface with task-specific textual supervision.

C.2 Latent Feature Organization and Optimization

Pose-Aware Feature Compilation. Before the temporal learner can reason over sequences, we enforce a disciplined organization of the features produced by the Vid2Pose encoder. Every embedding $z_t \in \mathbb{R}^{2048}$ is accompanied by its gloss supervision signal g_t and confidence mask m_t, giving us a tuple $\langle z_t, g_t, m_t \rangle$. As shown in Fig. 7, this tuple is passed through a *transformer stack* that performs (i) variance-preserving layer normalization, and (ii) stochastic frame dropping (with probability ρ) to avoid overfitting to repetitive segments. The process aligns pose cues with textual semantics via the *GlossAlign* function formalized in Algorithm 1. GlossAlign assigns each feature to the gloss interval containing its temporal center; features without an overlapping annotated interval are ignored by the confidence mask.

Algorithm 1 PoseAwareFeatureCompilation

procedure COMPILE($\{z_t, g_t, m_t\}_{t=1}^{T}$)
  $z_t \leftarrow \text{LayerNorm}(z_t)$
  if $\text{Bernoulli}(\rho) = 1$ **then**
    $z_t \leftarrow \mathbf{0}; m_t \leftarrow 0$ ▷ Stochastic frame drop
  end if
  $a_t \leftarrow \text{GlossAlign}(z_t, g_t)$
  return $\{(z_t, a_t, m_t)\}_{t=1}^{T}$
end procedure

Gloss-Aware Latent Distillation. With the compiled tuples we run a *knowledge distillation* objective that pushes the latent vectors toward a compact space guided by gloss embeddings. A teacher backbone f_{ResNet} (based on ResNet-34) produces logits y_t^{teacher}. The latent vectors generate student logits y_t^{student} through a two-layer perceptron. The KL divergence between both streams, weighted by gloss frequency priors, ensures that rare glosses receive more attention. This formulation adapts *SeqKD* to our latent pose space. Algorithm 2 presents the complete hierarchical training procedure.

Adaptive Feature Pruning. To optimize the information density of the latent space, we monitor each latent dimension using its activation variance. Specifically, dimensions whose activation variance fails to surpass a dynamic threshold τ are identified as redundant and subsequently masked:

$$\text{Prune } z^{(d)} \text{ if } Var(z^{(d)}) < \tau \tag{15}$$

This mechanism reduces the effective latent width to approximately 1024 and captures relationships among the pose formats. A masking operation preserves structural compatibility and the tensor shapes expected by the subsequent BiLSTM and Transformer layers:

$$\hat{z}_t = z_t \odot M, \quad M_d = \begin{cases} 1 & \text{if } Var(z^{(d)}) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

The mask is estimated from training features after latent-space training and then fixed during downstream adaptation.

Multi-Scale Feature Augmentation. To combat data scarcity in sign language datasets, we apply a $N = 10$-fold temporal augmentation during feature extraction Zhang et al. (2023a). For each video sample i, the SMKD model Zhang et al. (2023a) generates N augmented versions through (i) random temporal scaling $\sigma \in [0.8, 1.2]$, (ii) spatial jittering with probability $p_{jit} = 0.3$, and (iii) feature-space Gaussian noise $\mathcal{N}(0, 0.01)$. The augmented features are stored in HDF5 format with keys $\{i\}_j$ where $j \in [0, N-1]$ indexes the augmentation version. This representation supports efficient random access, expands the effective data size from M to $M \times N$, and applies augmentation directly to cached features with low computational overhead.

Algorithm 2 HierarchicalLatentTraining

procedure TRAINLATENT($\mathcal{D}$)
  for epoch $= 1 \ldots E$ **do**
    for batch $\in \mathcal{D}$ **do**
      $\{(z_t, a_t, m_t)\} \leftarrow \text{COMPILE}(\text{batch})$
      $y^{\text{teacher}} \leftarrow f_{\text{ResNet}}(\text{batch})$
      $y^{\text{student}} \leftarrow f_{\text{MLP}}(\{z_t\})$
      $\mathcal{L}_{\text{KD}} \leftarrow \text{KL}(y^{\text{student}}, y^{\text{teacher}})$
      $\mathcal{L}_{\text{CTC}} \leftarrow \text{CTC}(y^{\text{student}}, \text{gloss})$
      $\mathcal{L} \leftarrow \lambda_{\text{KD}}\mathcal{L}_{\text{KD}} + \lambda_{\text{CTC}}\mathcal{L}_{\text{CTC}}$
      Update params via Adam
    end for
    Apply the fixed training-feature mask M to the latent sequence
  end for
end procedure

C.3 Latent-Space Temporal Modeling and Decoding

Hybrid Temporal Modeling. The distilled latent sequences feed a hybrid backbone that remains modality-agnostic. First, a *TemporalConv stack* (two K5 blocks interleaved with stride-2 pooling) compresses sequences from T frames to $T' \approx T/4$, producing features $h_t \in \mathbb{R}^{1024}$. These features then pass through a *bidirectional LSTM* (2 layers, hidden size 512 per direction) to capture long-range dependencies, yielding $u_t \in \mathbb{R}^{1024}$. Because all computations remain within the latent space, we avoid heavy pixel-level operations during inference.

Algorithm 3 LatentSequenceDecoding

procedure DECODE($H=\{u_t\}_{t=1}^{T'}$, beam B)
  Init $\mathcal{B} \leftarrow \{\langle \text{BOS}, 0\rangle\}$
  for $t = 1 \dots T_{\max}$ **do**
    $\mathcal{B}' \leftarrow \emptyset$
    for hyp $\langle y, s\rangle \in \mathcal{B}$ **do**
      $p(\cdot) \leftarrow \text{DecoderStep}(y, H)$
      for w in top-k of p **do**
        $s' \leftarrow s + \log p(w)$
        $\mathcal{B}' \leftarrow \mathcal{B}' \cup \{\langle y \oplus w, s'\rangle\}$
      end for
    end for
    $\mathcal{B} \leftarrow$ top-B of $\mathcal{B}'$
    if all beams end with EOS **then break**
    end if
  end for
  return best hyp in $\mathcal{B}$
end procedure

CTC-Constrained Transformer Refinement. A *Transformer encoder-decoder* (6 layers each, width 256) refines the BiLSTM outputs into gloss sequences. The encoder ingests u_t after projection to 256 dimensions. The decoder consumes BPE-tokenized gloss strings. We attach a *CTC head* directly to the encoder outputs to enforce monotonic alignment, and the architecture accepts the latent tensors in their existing shape. The joint objective combines sequence-level cross-entropy, encoder-side CTC, and a *latent-space regularizer* that encourages successive encoder states to remain within a Lipschitz ball.

Beam Search with Latent Feedback. During inference we run a 5-beam search over the gloss decoder, using the same beam configuration as the training recipe. The final gloss output uses the encoded RGB features directly. Algorithm 3 details the beam search decoding procedure.

Training Optimization and Regularization. To stabilize convergence and improve generalization, we employ a suite of optimization techniques. The learning rate follows a Noam scheduler Vaswani et al. (2017) with warmup steps $W = 4000$:

$$\eta(t) = d_{model}^{-0.5} \cdot \min(t^{-0.5}, t \cdot W^{-1.5}) \tag{17}$$

where $d_{model} = 256$. We apply *differentiated dropout* with rates tailored to each component: $p_{attn} = 0.3$ for attention weights, $p_{relu} = 0.5$ for activation layers, and $p_{res} = 0.4$ for residual connections, preventing over-reliance on any single pathway. Label smoothing with $\epsilon = 0.1$ regularizes the output distribution. Finally, we adopt *checkpoint averaging* over the top-5 models ranked by validation WER.

D Temporal Compression and Restoration Analysis

D.1 Temporal Compression Analysis

Figure 8 aligns gloss outputs with SignXray feature indices and original video frames. One encoded feature typically aggregates approximately 5–10 frames, so a gloss is often supported by only one or two feature positions. In the highlighted case, Feature 6 contributes to both *IX* and *NOT-YET*, demonstrating that the representation retains multiple locally co-occurring sign cues within a compact feature. Appendix Sec. D.2 further analyzes how the downstream sequence model restores output order.

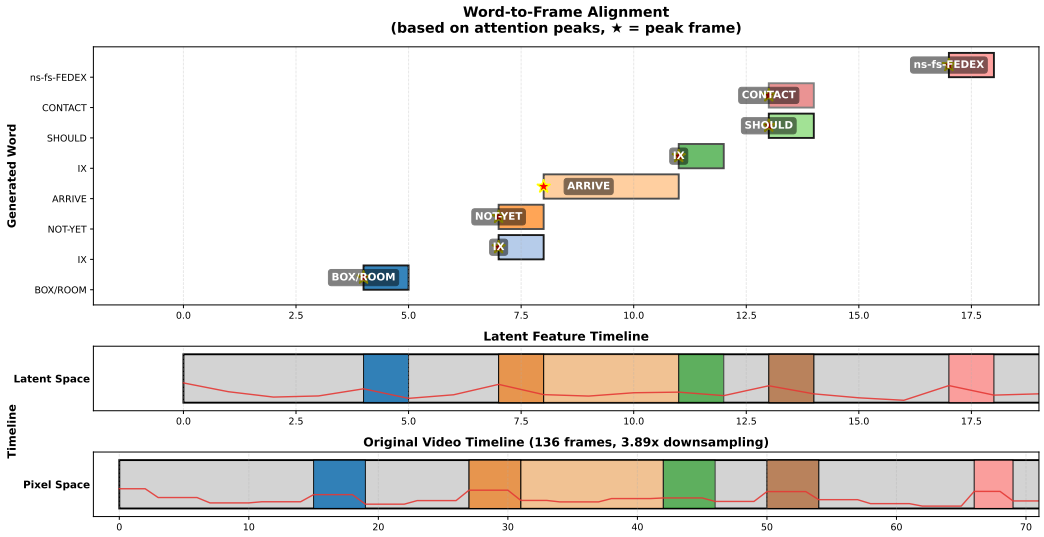


Figure 8: **Temporal compression in the SignXray representation.** We align output glosses, encoded feature positions, and original RGB frames. Each encoded feature summarizes approximately 5–10 input frames, and the red trace visualizes relative attention intensity. A single high-response feature can support more than one temporally adjacent gloss, illustrating the information density of the learned representation.

D.2 Downstream Temporal Restoration Analysis

The main paper shows that a single SignXray feature can retain cues from multiple adjacent signs. Here, we further examine whether temporal order is already explicit in the compressed representation or is restored by the downstream sequence model.

Figure 9 shows a sample whose strongest feature responses occur near the end of the video. In this case, the model emits three glosses after processing the full sequence. This pattern highlights two behaviors:

1. **Delayed Feature Consolidation:** Video2Pose assigns high responses after additional temporal context becomes available. The rising attention near the end indicates that the encoder consolidates evidence across the full clip.
2. **Contextual Restoration:** Once our Latent CSLR obtains these features, it restores them to their correct temporal positions. **For example**, the final output of the model is "#IF FRIEND GROUP/TOGETHER DEPART PARTY IX-1p JOIN IX-1p", and the reference is "#IF FRIEND GROUP/TOGETHER GO-OUT PARTY IX-1p JOIN IX-1p". The feature "PARTY," which is determined late in the sequence, returns to its correct output position. The remaining lexical error between "DEPART" and "GO-OUT" reflects the current model's resolution of closely related glosses.

This example suggests that the downstream sequence model learns to realign delayed feature evidence and recover the output chronology. Together with the main-paper analysis, it shows that compact SignXray features can preserve local sign cues and support temporal restoration by the task head.

E More Experimental Results

E.1 Evaluation Setup and Metrics

Evaluation scope. The core benchmarks comprise ASLLRP Neidle & Metaxas (2025) (1,686/210/212 train/development/test samples), PHOENIX2014-T Camgöz et al. (2018) (7,096/519/642), CSL-Daily Zhou et al. (2021b) (18,401/1,077/1,176), and WLASL-2000 Li et al.

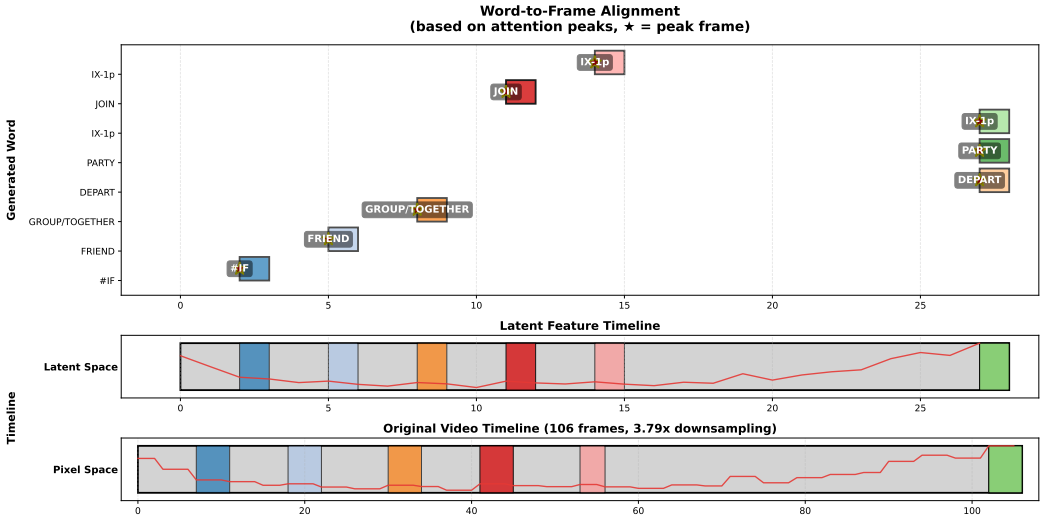


Figure 9: **Further analysis of the reasons for excellence.** To verify the entire pipeline of our model for the CSLR process and understand how its mechanism operates, we selected a sample in the translation process where all video features were concentrated at the relatively tail end. In this sample, the last video feature encoded information about three postures.

(2020a) (8,618/479/478). We further evaluate large-scale and multilingual transfer on OpenASL Shi et al. (2022), How2Sign Duarte et al. (2021), and three WMT23 SignSuisse languages through Uni-Sign and YouTube-SL-25 configurations Li et al. (2025); Tanzer & Zhang (2025). The evaluation uses BLEU Papineni et al. (2002), ROUGE, BLEURT, WER Klakow & Peters (2002), and P-I accuracy Li et al. (2020b). All preprocessing, teacher-feature generation, vocabulary construction, normalization statistics, and encoder adaptation follow the corresponding official train/development/test partition; held-out videos and labels are excluded from these operations. Each reported configuration is trained with ten random seeds; the checkpoint with the best development metric is selected for the test evaluation.

For sign recognition, we measure:

1. Word Error Rate (WER): As the standard metric for Continuous Sign Language Recognition (CSLR), WER quantifies the alignment between the predicted gloss sequence and the ground truth by calculating the normalized edit distance:

$$\text{WER} = \frac{S + D + I}{N} \tag{18}$$

where S, D, and I represent the number of *substitutions*, *deletions*, and *insertions* required to transform the hypothesis into the reference, respectively. N denotes the total number of glosses in the reference. A lower WER indicates higher recognition stability, particularly in capturing the subtle transitions between sign gestures.

2. BLEU-n Score: To evaluate the translation quality of the generated glosses or sentences, we employ the BLEU score, which measures the n-gram precision between the candidate and reference sequences:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{k} w_n \ln p_n\right) \tag{19}$$

Here, p_n denotes the modified n-gram precision, and w_n represents the uniform weight $1/k$. The Brevity Penalty (BP) is incorporated to penalize overly short predictions:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{20}$$

where c is the length of the candidate and r is the reference length. Higher BLEU scores reflect better semantic consistency and linguistic fluency of the SignXray decoder.

ASLLRP											
	Method	Dev					Test				
		Rouge	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	BLEU1	BLEU2	BLEU3	BLEU4
Sign2Gloss	C2SLR Zuo & Mak (2022)	52.50	50.10	38.20	29.50	24.76	52.30	49.95	38.00	29.35	24.58
	C2SLR + SignXray	**56.49**	**55.01**	**40.37**	**31.77**	**25.23**	**56.57**	**54.16**	**41.68**	**30.63**	**25.41**
	CorrNet+ Hu et al. (2024)	53.80	51.40	39.50	31.20	26.14	53.60	51.25	39.30	31.05	25.96
	CorrNet+ + SignXray	**59.47**	**53.79**	**40.39**	**33.56**	**28.54**	**59.73**	**55.58**	**41.99**	**34.51**	**28.38**
	SLTUNET Zhang et al. (2023a)	54.20	52.10	40.10	31.80	26.48	54.05	51.90	39.95	31.65	26.31
	SLTUNET + SignXray	**56.65**	**51.84**	**42.48**	**35.35**	**29.80**	**56.48**	**51.65**	**42.31**	**35.19**	**29.64**

Table 8: **Controlled plug-and-play comparison on ASLLRP** Neidle et al. (2022a); Neidle & Metaxas (2025). Each +SignXray configuration replaces the original visual encoder while retaining the corresponding downstream architecture, data split, training protocol, and decoding procedure. The paired models use the same RGB video source at deployment, with SignXray producing the compact feature interface.

Method	Modality		Test					
	Pose	RGB	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEURT
Gloss-free sign language translation								
GloFE-VN Lin et al. (2023)	✓		14.9	7.3	3.9	2.2	12.6	31.7
YouTube-ASL Uthus et al. (2024)	✓		37.8	24.1	16.9	12.4	–	46.6
MSLU Zhou et al. (2025)	✓		20.1	7.7	–	2.4	17.2	–
SLT-IV Tarrés et al. (2023)		✓	34.0	19.3	12.2	8.0	–	–
C^2RL Chen et al. (2024a)		✓	29.1	18.6	12.9	9.4	27.0	–
FLa-LLM Chen et al. (2024b)		✓	29.8	19.0	13.3	9.7	27.8	–
VSLT (gloss-free) Zhao et al. (2026)		✓	39.0	25.5	18.8	13.4	33.6	–
SignLlama (gloss-free) Gan et al. (2026a)		✓	30.5	20.0	16.9	9.9	30.3	–
SignMusketeers Gueuwou et al. (2025a)		✓	41.5	27.2	19.3	14.3	–	–
SSVP-SLT Rust et al. (2024)		✓	43.2	28.8	20.8	15.5	38.4	49.6
Uni-Sign Li et al. (2025)	✓		40.4	26.8	19.3	14.5	34.3	48.6
Uni-Sign Li et al. (2025)	✓	✓	40.2	27.1	19.7	14.9	36.0	49.4
Uni-Sign + SignXray	✓	✓	**43.5**	**29.5**	**21.6**	**16.1**	**38.7**	**54.9**

Table 9: **Gloss-free SLT results on How2Sign.** Published results are reported from the original studies; Uni-Sign baselines are reproduced from Uni-Sign Li et al. (2025). For Uni-Sign, modality marks follow its original training convention; the SignXray encoder receives only RGB video at inference.

Configuration	American Sign Language (ase) How2Sign	sgg WMT23 SignSuisse	ssr WMT23 SignSuisse	slf WMT23 SignSuisse
None	– / 31.2	– / 26.2	– / 7.2	– / 22.8
YouTube-SL-25 (ASL)	30.2 / 46.6	9.0 / 27.9	5.4 / 11.1	9.3 / 20.8
YouTube-SL-25 + SignXray (ASL)	**– / 49.3**	**– / 28.8**	**– / 11.4**	**– / 22.5**
YouTube-SL-25 (Full)	29.7 / 47.9	9.5 / 37.7	6.8 / 18.8	9.3 / 25.2
YouTube-SL-25 + SignXray (Full)	**– / 52.5**	**– / 38.6**	**– / 20.0**	**– / 26.9**

Table 10: **YouTube-SL-25 baselines for sentence-level sign language translation** Tanzer & Zhang (2025). Translation is measured with BLEURT; each cell reports zero-shot / fine-tuned performance. Published baselines reproduce the translation portion of Table 2 in the original paper. The paired +SignXray rows retain the corresponding fine-tuning setting and do not report zero-shot evaluation.

3. Per-Instance (P-I) Top-1 Accuracy: Per-instance (P-I) accuracy is the primary metric used to evaluate the overall recognition performance of **SignXray** across the entire test set. It treats each video sample as an independent and equal evaluation unit, representing the global classification accuracy (Micro-average). Formally, it is defined as:

$$\text{Acc}_{P\text{-}I} = \frac{\sum_{i=1}^{C} n_{i,i}}{\sum_{i=1}^{C} N_i} \tag{21}$$

where C denotes the total number of sign categories in the dataset (e.g., 2,000 for WLASL2000), $n_{i,i}$ is the number of samples from class i correctly identified by the model, and N_i is the total number of test samples belonging to category i.

E.2 Additional Benchmark Results

Table 8 provides a controlled test of SignXray across three established downstream architectures on ASLLRP. Each pair preserves the downstream task head, data split, optimization recipe, and decoding procedure; the visual encoder interface is the experimental replacement. With C2SLR, SignXray

System	Encoder integration	Test BLEU-4 ↑			
		PHOENIX	CSL-Daily	How2Sign	OpenASL
Concurrent SLT systems: baseline / SignXray extension					
VTaMo Hu et al. (2026)	Frozen CLIP visual interface	28.86	27.16	18.47	25.94
VTaMo + SignXray	SignXray visual interface	**30.84**	**27.59**	**19.72**	**28.06**
SignThought Jiang et al. (2026)	Pre-extracted visual feature interface	27.22	23.92	13.39	19.55
SignThought + SignXray	SignXray visual interface	**30.12**	**24.47**	**13.87**	**20.22**
SIGNET Asasi et al. (2026)	Pose-expert visual interface	27.82	28.51	15.40	22.77
SIGNET + SignXray	SignXray RGB expert/interface	**28.33**	**29.69**	**16.80**	**23.23**
CV-SLT + BoostSLT Han et al. (2026)	CV-SLT visual interface	33.32	32.49	12.03	21.80
CV-SLT + SignXray + BoostSLT	SignXray visual interface	**36.79**	**35.82**	**13.26**	**22.69**

Table 11: **Plug-and-play extension to concurrent SLT systems.** Baseline rows use the test results reported in the original papers when available; metrics absent from an original study are reproduced using its official implementation. Each SignXray row preserves the corresponding split, task head, optimization recipe, and decoding configuration, with SignXray as the only encoder substitution.

raises development/test ROUGE by 3.99/4.27 points and BLEU-4 by 0.47/0.83 points. The gains become particularly pronounced with CorrNet+: ROUGE increases by 5.67/6.13 points and BLEU-4 by 2.40/2.42 points. The SLTUNET pairing adds 2.45/2.43 ROUGE points and 3.32/3.33 BLEU-4 points. These repeated improvements across three structurally different systems demonstrate that the pose-rich representation transfers reliably and supplies a consistently stronger visual interface.

The How2Sign results in Table 9 further highlight the quality of the learned encoder. Recent RGB-only gloss-free VSLT and SignLlama baselines are also included for an up-to-date comparison. Adding SignXray to pose-and-RGB Uni-Sign improves all six metrics: BLEU-1/2/3/4 rise from 40.2/27.1/19.7/14.9 to 43.5/29.5/21.6/16.1, ROUGE increases from 36.0 to 38.7, and BLEURT increases substantially from 49.4 to 54.9. The enhanced model achieves the top performance across every reported metric in this comparison, showing that SignXray improves both lexical overlap and learned semantic quality on open-domain American Sign Language translation.

Table 10 expands this evidence to multilingual YouTube-SL-25 settings. Under ASL-only pretraining, SignXray improves fine-tuned BLEURT by 2.7 points on How2Sign and by 0.9, 0.3, and 1.7 points on the three SignSuisse languages. Under full multilingual pretraining, the respective gains are 4.6, 0.9, 1.2, and 1.7 points, reaching 52.5 BLEURT on How2Sign and 38.6/20.0/26.9 on `sgg`/`ssr`/`slf`. The consistent improvements across pretraining scales and four sign languages provide strong evidence that SignXray preserves transferable linguistic cues beyond a single dataset or language.

Table 11 completes the paired evaluation with four recent SLT systems. Baseline rows use original-paper results when available, and we reproduce unreported metrics with the corresponding official implementation. Each adjacent SignXray row preserves the dataset split, task head, optimization recipe, and decoding configuration; the visual encoder interface is the sole experimental substitution. SignXray improves every evaluated setting: VTaMo gains 1.98/0.43/1.25/2.12 BLEU-4 across PHOENIX, CSL-Daily, How2Sign, and OpenASL; SignThought gains 2.90/0.55/0.48/0.67; SIGNET gains 0.51/1.18/1.40/0.46; and the BoostSLT pairing gains 3.47/3.33/1.23/0.89. These results show that the unified face-hand-body representation strengthens recent alignment, latent-reasoning, multi-expert, and semantic-refinement systems without altering their downstream designs.

Together, Tables 8–10 and Table 11 validate the central promise of SignXray as a powerful plug-and-play sign encoder. Its gains persist across recognition and translation architectures, American and European sign languages, controlled and open-domain benchmarks, and both n-gram and learned semantic metrics. This breadth of improvement establishes the unified face-hand-body multi-prior space as an effective and broadly reusable representation for sign language understanding.

F DISCUSSION

SignXray learns a compact video representation from complementary face, hand, body, depth, and segmentation cues. These cues construct the training target, and the deployed encoder maps RGB videos directly to the resulting representation. The shared interface connects naturally to recognition and translation heads and enables efficient deployment with learned multi-prior knowledge.

For each paired experiment, we reproduce the corresponding downstream method with its original architecture, data split, optimization recipe, and decoding procedure, then replace only its visual encoder with SignXray. Improvements observed across these pairs therefore measure the encoder substitution while the task head remains fixed. All experiments use the standard train/development/test partitions of each dataset; teacher features, vocabulary and normalization statistics are derived from the training partition, and the resulting pipelines were checked for cross-partition leakage. For every configuration, ten independent random seeds are used during development, and development performance selects the test checkpoint. The selected configuration is evaluated five times on the test split, and every table reports the resulting mean. Each baseline–SignXray pair follows this same selection-and-five-run evaluation protocol, so the reported gains reflect matched test averages from repeated evaluation. For the PHOENIX2014-T paired SLT evaluation in Table 3, the ten independent development-seed evaluations have a standard deviation below 0.61% under the fixed training protocol. A paired bootstrap analysis of this PHOENIX2014-T evaluation yields a mean gain of $+3.27$ with a 95% confidence interval of $[+1.14, +3.87]$, excluding zero. The recurring gains across architectures, datasets, tasks, and languages demonstrate the reliable value of the SignXray encoder. The efficiency measurements report the measured runtime and power of the complete inference configurations listed in Table 6, measured on an eight-GPU NVIDIA Quadro RTX 8000 system (48GB per GPU; driver 560.28.03). The FPS ratios describe the tested RGB-to-task and SignXray-to-task pipelines on ASLLRP and establish a concrete deployment advantage for the compact representation.

The stage boundaries are explicit. Stage 1 updates the pose-space encoder and its semantic decoder with the text-generation, word-matching, and contrastive objectives; it produces the unified teacher features. Stage 2 freezes this teacher and updates only the RGB Video2Pose module with feature regression. The downstream CSLR and translation losses, including KD, CTC, and sequence-level decoding, are applied after the encoder interface has been formed and are not additional Stage-2 objectives. In each paired comparison, the original task head, loss definitions, decoder, and checkpoint protocol are retained; the visual encoder output is replaced at the same feature interface, with any projection required by the original head kept unchanged. This clear division preserves the contribution of the multi-prior encoder across diverse downstream systems.

The plug-and-play interface is designed for RGB deployment. Pose files, semantic labels, and cached HDF5 features construct or adapt the encoder during training, and the deployed path processes a video as $Z = E_{\mathrm{SignXray}}(V)$ before passing it to the selected task head. The reported efficiency boundary includes the RGB encoder, the selected temporal head, and decoding for the listed configuration, while teacher extraction is an offline training cost. This separation delivers a practical encoder that retains rich human priors within a single RGB pathway.

For the two-stage construction, Stage 1 learns the unified teacher space from the five prior streams and Stage 2 freezes that space while optimizing the RGB Video2Pose module against its aligned feature grid. The feature projection preserves the fixed 2,048-dimensional interface, and variance masking selects active channels while maintaining the tensor shape expected by downstream heads. Stage 1 uses the dataset word-index codebook for its text, word-matching, and contrastive objectives; this semantic decoder supplies training supervision, while the RGB encoder carries the learned representation to deployment. Temporal projection aligns the RGB sequence to the teacher feature grid before the feature loss is evaluated, with each compact feature covering the local window described in Sec. D.2. The unified teacher is learned from the ASLLRP training split. For transfer experiments, the resulting encoder is adapted with the same official training split used by the corresponding downstream method, and development and test videos remain evaluation-only. This construction turns structured semantic supervision into a compact and effective visual interface.

The cross-dataset rows follow the same rule. OpenASL, How2Sign, PHOENIX2014-T, CSL-Daily, and the YouTube-SL-25 configurations use their official target-data partitions for adaptation and evaluation; external pretraining data are restricted to the corpus named in the corresponding config-

uration. Vocabulary, normalization, feature masks, and any target-data adapter are estimated from the training partition of that configuration. Teacher fitting and adapter estimation use only that training partition, and the RGB encoder receives the same source video used by its paired baseline. This consistent protocol supports strong, comparable performance across languages and downstream settings.

The representation dimensions are fixed before downstream training. The five per-frame records have widths 384, 258, 165, 576, and 576, which concatenate to $1{,}959$ channels. Source-specific encoders preserve the temporal index, the fusion layer operates on this per-frame record, and PAD-MATCH applies its learned projection to produce $T \times 2{,}048$ tokens. The ViT processes these tokens on the shared temporal grid; any $T \to T/4$ reduction belongs to the selected temporal head. The variance mask selects active channels elementwise while preserving the external 2,048-dimensional interface, so every paired head consumes the same declared feature shape. The fixed interface is a key strength that makes SignXray broadly compatible with established temporal models.

At the technical level, the experiments establish a clear separation between representation construction, RGB transfer, and task-specific use. The five prior streams define the teacher target; the RGB Video2Pose module learns the fixed 2,048-dimensional interface; and each downstream head consumes that interface under its corresponding data split and evaluation protocol. The paper makes the data partitions, feature dimensions, loss boundaries, deployment path, paired replacement protocol, and efficiency scope explicit. Implementation-level choices naturally follow the selected downstream system and hardware and can be verified and extended through the released code and configuration files. This transparent modular design strengthens the practical and scientific value of SignXray.

Taken together with the ablations and additional benchmark results, these experiments show that multi-prior supervision can be consolidated into a compact RGB representation and transferred to several sign-language tasks and datasets. Training-time prior extraction and downstream task modeling remain separate from the reusable encoder interface, allowing each component to be developed independently. Together, this combination establishes SignXray as a powerful, practical, and broadly reusable foundation for next-generation sign language understanding.