SignVerse-2M: A Two-Million-Clip Pose-Native Universe of 25+ Sign Languages

Sen Fang[1], **Hongbin Zhong**[2], **Yanxin Zhang**[3]
Dimitris N. Metaxas[1]
[1]Rutgers University [2]Georgia Institute of Technology [3]NVIDIA

Dataset Project Page Benchmark

Abstract

Existing large-scale sign language resources typically provide supervision only at the level of raw video-text alignment and are often produced in laboratory settings. While such resources are important for semantic understanding, they are not naturally suited to open-world recognition and translation, or to modern pose-driven sign language video generation frameworks: 1. RGB-based pretrained recognition models depend heavily on fixed backgrounds or clothing conditions during recording, and are less robust in open-world settings than style-agnostic pose-processing models. 2. Recent pose-guided image/video generation models mostly use a unified keypoint representation such as DWPose as their control interface. At present, the sign language field still lacks a data resource that can directly interface with this modern pose-native paradigm while also targeting real-world open scenarios.

We present **SignVerse-2M**, a large-scale multilingual pose-native dataset for sign language pose modeling and evaluation. Built from publicly available multilingual sign language video resources, it applies DWPose in a unified preprocessing pipeline to convert raw videos into 2D pose sequences that can be used directly for modeling, resulting in a consolidated corpus of about two million clips covering more than 25 sign languages. Unlike many laboratory datasets, this resource retains the shooting conditions and speaker diversity in the real world, excluding background changes and clothing differences. Therefore, it is more suitable for researching robust sign language models in open scenarios. Toward this goal, we further provide the data construction pipeline, task definitions, and a simple SignDW Transformer baseline, demonstrating the feasibility of this resource for multilingual pose-space modeling and discussing the evaluation claims it can support as well as its current limitations.

1 Introduction

In recent years, the scale and coverage of sign language datasets have grown substantially with the expansion of online video resources and multimodal learning methods Duarte et al. [2021b], Camgöz et al. [2018], Cihan Camgöz et al. [2020], Forster et al. [2012], Joze and Koller [2018], Li et al. [2020]. However, existing large-scale sign language resources still face two main limitations: first, they are mostly designed for tasks such as sign language recognition, retrieval, or sign-to-text translation Hu et al. [2023], Jiang et al. [2021], Tarrés et al. [2023], Cihan Camgöz et al. [2020], Koller [2020], with supervision centered on weak or strong alignment between videos and text Camgöz et al. [2018], Duarte et al. [2021b], Albanie et al. [2020]; second, they often involve only a small number of signers

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

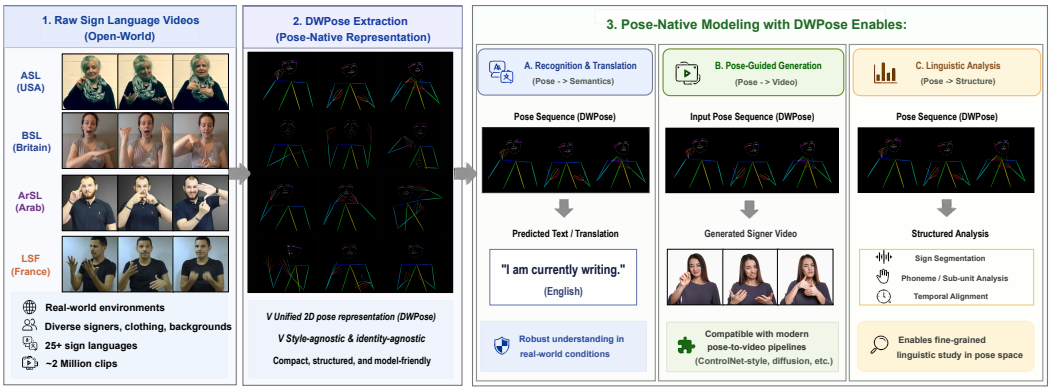

Figure 1: **Overview of SignVerse-2M.** SignVerse-2M organizes large-scale public sign language videos into a unified pose-native interface for multilingual sign language modeling. This representation is directly compatible with modern pose-driven generation pipelines and can serve as an intermediate control space for digital human or avatar generation Chen et al. [2023], Cai et al. [2023], Zwitserlood et al. [2004], Zhang and Agrawala [2023].

recorded with fixed clothing and backgrounds, in environments that are neither dynamic nor diverse and are overly idealized Forster et al. [2012, 2014], Von Agris and Kraiss [2010], Boháček and Hrúz [2022]. These resources are valuable for understanding what is being expressed in sign language videos, but they still lack suitable data representations and evaluation interfaces for problems that demand strong robustness in real-world recognition or in generation conditioned on complex textual inputs Fang et al. [2025c,a,d]. This issue is especially pronounced in multilingual sign language settings, where raw videos contain strong nuisance factors such as background, viewpoint, signer identity, and recording conditions, while different datasets often adopt **incompatible representations, preprocessing pipelines, and evaluation protocols**, making model comparison and interpretation of conclusions difficult Yin et al. [2022], Rayner et al. [2016], Moryossef and Müller [2021].

This leads us to believe that what sign language research truly lacks today is not simply more videos, but rather a **unified, pose-driven evaluation resource for in-the-wild sign language tasks** Saunders et al. [2020b], Fang et al. [2025a], Baltatzis et al. [2024]. This need is particularly evident in sign language generation: existing models based on Mesh Human, MediaPipe, or non-standard OpenPose derivatives Cai et al. [2023], Lugaresi et al. [2019], Cao et al. [2017] are difficult to develop further with strong modern backbones released by major companies because they are not compatible with contemporary pose-to-video pipelines Ma et al. [2024], Zhang et al. [2025], Cheng et al. [2025].

> *Over the past three years, a large number of pose-driven human image and video generation models have gradually converged on a unified technical interface: they first extract standardized keypoints such as DWPose Yang et al. [2023] from real-world videos, and then use these keypoints to drive ControlNet-style control modules Zhang and Agrawala [2023], Peng et al. [2025] or other pose-conditioned generation frameworks Zhang et al. [2023], Chan et al. [2019]. In other words, in the broader visual generation community, DWPose Yang et al. [2023], as a high-accuracy and high-speed 2D skeleton estimation model, has effectively become one of the* **standard input representations for modern pose-driven generation**.

In contrast, although the sign language field now has access to more and more video resources, it still lacks a standardized data interface that is directly compatible with this modern generative paradigm Stoll et al. [2020], Saunders et al. [2020b], Fang et al. [2025a]. This is particularly important for the NeurIPS 2026 Evaluations & Datasets Track: the value of a dataset should not be measured only by scale, but also by whether it supports clear, comparable, and interpretable evaluation.

Motivated by this perspective, we propose **SignVerse-2M**. Rather than releasing another raw video-text parallel corpus, we build on existing large-scale multilingual sign language video resources and apply DWPose for unified pose extraction and standardized processing, converting them into a large-scale pose-native sign language representation dataset covering more than 25 sign languages and about two million clips. As shown in Figure 1, our goal is not to recreate yet another video-

Dataset	Duration(h)	Vocabulary(k)	Annotation Type	Year	Domain
KETI Ko et al. [2019]	27.99	0.49	Spoken Text	2019	Emergency situations
PHOENIX-2014T Camgoz et al. [2018]	10.50	3.90	Spoken Text, Gloss	2018	Weather Forecast
CSL Daily Zhou et al. [2021]	23.27	4.60	Spoken Text, Gloss	2021	Daily life
OpenASL Shi et al. [2022]	288	33	Spoken Text	2022	Youtube (news + vlogs)
How2Sign Duarte et al. [2021a]	79.10	22.40	Spoken Text	2021	Instructional
ASLLRP Neidle et al. [2022]	80	2.1	Gloss	2022	Comprehensive
YouTube-ASL Uthus et al. [2023]	∼1000	60	Spoken Text	2023	Youtube (open-domain ASL)
YouTube-SL-25 Tanzer et al. [2024]	3207	∼374	Spoken Text	2024	Youtube (25+ languages)
SignVerse-2M (ours)	3207	∼374	Spoken Text, Pose	2026	Pose-native open-world multilingual

Table 1: **Comparison with representative and widely used sign language datasets:** We compare SignVerse-2M with representative and widely used sign language datasets under a unified layout, covering annotation type, scale, and domain. The top five rows mainly correspond to spoken-text translation benchmarks, while ASLLRP, YouTube-ASL, and YouTube-SL-25 place our resource in broader and more open-domain settings. Duration and vocabulary are reported as corpus-level statistics; "-" denotes values that are unavailable or not directly comparable, and approximate values are marked with "∼".

text corpus, but to construct a data interface that can directly support sign language generation research: 1. it reduces variation in visual appearance and shifts the focus back to the hand, body, and upper-body motions themselves; 2. it provides a more natural input-output space for cross-lingual modeling, unified training, and standardized evaluation; 3. more importantly, this representation is naturally compatible with existing pose-to-pose, pose-to-video, and other pose-driven generation pipelines, allowing sign language generation to move beyond starting directly from pixel space and instead build on mainstream pose-conditioned modeling paradigms, benefiting from large-scale video models developed by major companies. As shown in Table 1, representative existing datasets differ substantially in annotation type, scale, and domain, but most are still organized around video–text supervision rather than a unified pose-native interface.

To address these questions, this paper makes four main contributions:

- We construct **SignVerse-2M**, a large-scale, multilingual, uniformly represented pose-native sign language dataset that systematically converts public video resources into a pose-sequence corpus suitable for generative tasks.
- We explicitly position this resource as a **sign language interface for modern pose-driven generative models**, emphasizing its compatibility with the DWPose-centered control paradigm rather than treating pose extraction merely as a preprocessing byproduct.
- We introduce **task settings and a simple SignDW Transformer baseline** for sign language generation research to validate the practical usability of this resource and provide a reproducible starting point for future work.
- We discuss the **strengths** of this unified pose representation in terms of real-world robustness, cross-lingual transfer, and deployment feasibility, as well as its **limitations** in fine-grained hand information, non-manual features, and linguistic completeness.

In summary, SignVerse-2M is intended to support the shift of sign language research from idealized laboratory settings toward real-world pose modeling.

2 Background and Positioning

2.1 Traditional Video-Text Sign Language Corpora

The Development of Sign Language Data. Deep learning based sign language modeling typically requires large amounts of data, and in general, more data is better Camgöz et al. [2018], Cihan Camgöz et al. [2020], Hu et al. [2023]. From Phoenix Forster et al. [2012, 2014], MS-ASL Joze and Koller [2018], How2Sign Duarte et al. [2021b], Sign.MT, and Prompt2Sign Fang et al. [2025a] to YouTube-SL-25, the overall trend has been toward either larger datasets or broader language coverage Li et al. [2020], Albanie et al. [2020, 2021], Shi et al. [2022], Yin et al. [2022]. Most existing large-scale sign language datasets are built around the mapping between video and text, and are primarily intended for tasks such as sign language recognition, retrieval, and sign-to-text translation Cihan Camgöz et al. [2020], Tarrés et al. [2023], Jiang et al. [2021], Koller [2020]. For these tasks, raw video preserves

the richest visual information, while text provides clear semantic supervision, making video-text parallel corpora the natural data format. However, when the research goal shifts to translation in open environments or to modern sign language video generation Saunders et al. [2020b], Fang et al. [2025a], Baltatzis et al. [2024], Yin et al. [2024], the problem changes. The focus is no longer only on whether semantics are aligned, but also on whether long motion sequences are coherent, whether hand and upper-body movements are natural, whether representations are consistent across samples, and whether different models can be fairly compared under a unified interface Stoll et al. [2020], Saunders et al. [2020b], Walsh et al. [2025]. In this setting, directly using raw video often mixes irrelevant factors such as background, signer appearance, camera viewpoint, and recording conditions into both modeling and evaluation, thereby weakening the analysis of motion itself Boháček and Hrúz [2022], Fang et al. [2025d].

2.2 From Laboratory Distributions to Real-World Distributions

Why Choose a Pose-Native Representation? Pose representation provides a more suitable intermediate space for this problem. By uniformly mapping raw videos into keypoint sequences Yang et al. [2023], Cao et al. [2017], Lugaresi et al. [2019], Cai et al. [2023], models can focus more directly on the temporal structure and body geometry of sign language motion without having to simultaneously handle complex appearance variation Saunders et al. [2020b, 2021], Stoll et al. [2020]. This is especially important in multilingual settings, where the visual differences across languages, video sources, and signers are often much larger than the domain gaps commonly seen in text corpora Yin et al. [2022], Fang et al. [2025a]. More importantly, over the past three years, DWPose Yang et al. [2023] has become one of the standardized control interfaces for a large number of pose-driven image and video generation models Ma et al. [2024], Zhang et al. [2025], Cheng et al. [2025]. Whether in ControlNet-style conditional control Zhang and Agrawala [2023], Zhang et al. [2023], Peng et al. [2025] or in other generation frameworks that use human keypoints as intermediate conditions Chan et al. [2019], Saunders et al. [2020a], a stable, general, and large-scale extractable pose representation is essential. A concrete example is that RGB-based recognition models are difficult to apply directly in real-world environments: once the background, person, or clothing changes, the input video features can differ significantly Koller [2020], Boháček and Hrúz [2022]. Pose-based inputs or generation, in contrast, are more robust Fang et al. [2025c,b].

2.3 Positioning of SignVerse-2M

Relation to Existing Multilingual Sign Language Resources. SignVerse-2M inherits from existing public multilingual sign language video corpora Duarte et al. [2021b], Albanie et al. [2020, 2021], Shi et al. [2022], Yin et al. [2022], but its research objective is different. Prior work primarily asks how to collect, align, and organize cross-lingual sign language video-text data from the web in order to support translation tasks Cihan Camgöz et al. [2020], Camgöz et al. [2018]. SignVerse-2M instead focuses on a different question: when the target setting shifts toward the real world, can these public video resources be transformed into a unified pose space that serves as reusable, extensible, and comparable infrastructure for generative research Yang et al. [2023], Cheng et al. [2025], Zhang and Agrawala [2023]? We position SignVerse-2M as a pose-native resource not because pose extraction itself is especially complex, but because it enables sign language data to connect naturally with mainstream generative paradigms Ma et al. [2024], Zhang et al. [2025], Peebles and Xie [2023], making the definition of sign language motion representation and its associated evaluation protocol part of the scientific object itself rather than merely a preprocessing step Fang et al. [2025a]. We therefore hope that reviewers and readers will understand this work as a reconstruction of the *data representation, task interface, and evaluation object*, rather than as a simple expansion of existing video-text corpora.

Questions Addressed in This Paper. Based on the above considerations, this paper does not aim to prove that SignVerse-2M is superior to existing resources for every sign language task. Instead, it focuses on three more specific questions. First, can large-scale public videos be stably transformed into a unified pose corpus suitable for sign language generation and directly compatible with modern pose-driven generation frameworks Yang et al. [2023], Cheng et al. [2025], Ma et al. [2024]? Second, is this pose-space representation sufficient to support the training and comparison of multilingual generative models under real-world distributions Yin et al. [2022], Fang et al. [2025a,c]? Third, if used as an evaluation resource, what conclusions can it support, and where are its boundaries and

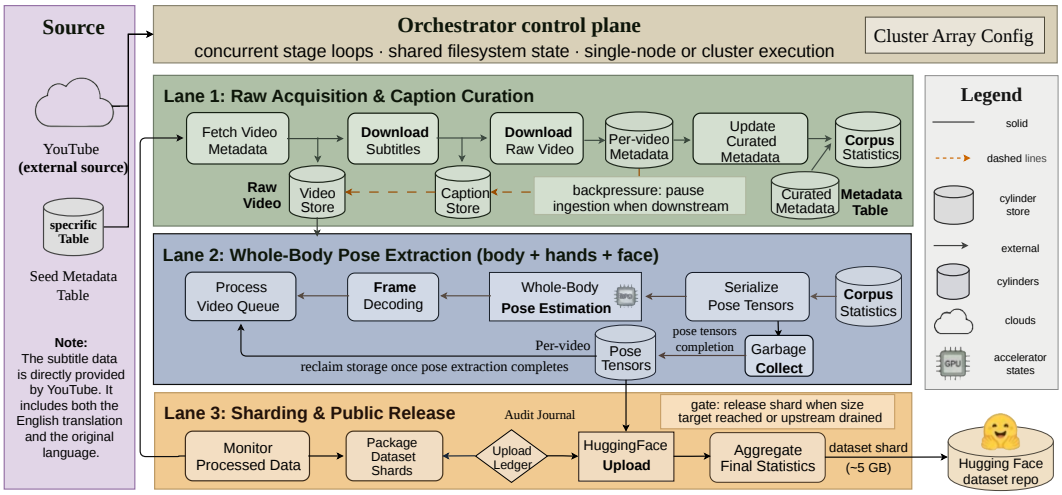

Figure 2: **Overview of the SignVerse-2M data processing pipeline.** The pipeline is organized into three main lanes: raw acquisition and caption curation, whole-body pose extraction, and sharding for public release. Starting from a manifest indexed by ‘video_id‘ and ‘sign_language‘, the system retrieves metadata, subtitles, and raw videos, converts each video into DWPose-based body, hand, and face keypoint sequences, and then packages the processed outputs into dataset shards for publication. The orchestration layer maintains status records, supports staged re-execution and failure recovery, and enables the corpus to be built incrementally rather than as a one-shot preprocessing export.

limitations Walsh et al. [2025], Koller [2020]? The remainder of the paper is organized around these three questions.

3 Data Infrastructure and Pipeline

3.1 Data Source and Curation Pipeline

Data Source. SignVerse-2M starts from publicly available multilingual sign language video resources and builds a large-scale corpus for pose-based modeling on top of them. We maintain a manifest centered on ‘video_id‘ and ‘sign_language‘, which drives the subsequent processing pipeline. The raw metadata table in the current repository contains 39,196 video entries and roughly 2M segments; each entry is fed into a traceable, resumable, and re-runnable processing pipeline with associated status records. Accordingly, the focus of SignVerse-2M is to build an infrastructure for automatic sign language video processing that turns open-world videos into a unified, trainable, and publishable pose-native corpus.

Pipeline Overview. Figure 2 summarizes the overall construction path of SignVerse-2M. The pipeline begins with manifest-driven sample management, proceeds through metadata and subtitle acquisition, and then moves into pose extraction, packaging, and publication. This layout is useful because it makes clear that the corpus is not defined by a single preprocessing script, but by a sequence of recoverable and independently executable stages.

Video Acquisition. Before pose extraction begins, we first complete video-level data acquisition. For each video ID in the manifest, the system retrieves the original platform metadata, downloads the raw video, and fetches the available subtitles in parallel. This stage continuously writes back fields such as the title, duration, subtitle languages, raw video path, processing time, and error status, allowing the metadata table to serve as both a sample inventory and a runtime record. As a result, the corpus construction process can proceed incrementally by entry and can resume from intermediate states after interruption.

Subtitle Structuring and Language Signals. Once subtitles are downloaded locally, they are converted into structured text objects instead of preserving the platform-exported format verbatim.

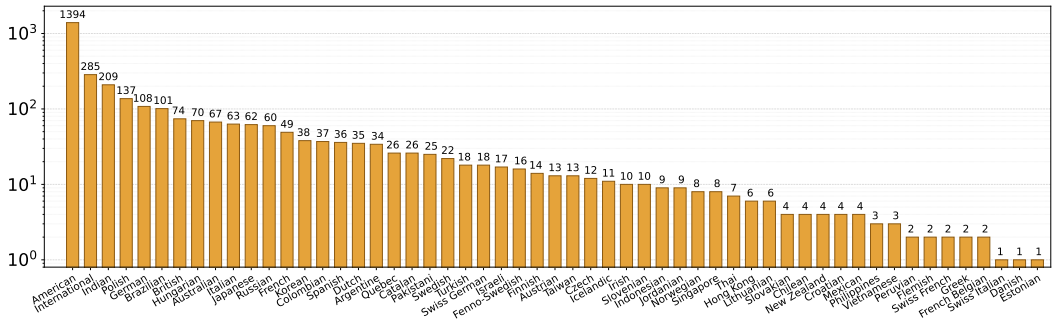

Figure 3: Distribution of content hours across sign languages in YouTube-SL-25. The x-axis lists sign language names, and the y-axis reports total hours on a logarithmic scale.

During processing, we remove WEBVTT control fields, timestamp lines, HTML tags, and zero-width characters, and apply rule-based normalization to repeated lines, broken punctuation, hyphen continuations, and ellipses. At the output level, we preserve both segment-level subtitles with start and end times and compact document-level text intended for training-time consumption. For English supervision, the system prioritizes native English subtitles; when no native English track exists, it automatically selects from available English translation variants and records the source language. Each video therefore receives a structured caption JSON containing multilingual subtitle text, temporal boundaries, and the provenance of the English supervision.

From Video-Text Data to Pose Data. After videos and subtitles have been acquired, the system converts each video into a pose sample. The current implementation decodes frames at 24 FPS and applies DWPose to extract body, hand, and facial keypoints from each frame. To support large-scale processing, the pipeline can read frames directly from the ‘ffmpeg‘ stdout stream instead of first materializing all frames as JPG files; a legacy path is retained for debugging and fallback. Pose outputs are aggregated per video into ‘poses.npz‘, which stores the video ID, sampling rate, total frame count, frame indices, frame dimensions, and per-frame keypoint payloads. After each video is processed, the system also writes a ‘.complete‘ marker and corresponding status records to distinguish unprocessed, running, and completed samples. Through this stage, open-world sign language videos are reorganized into a unified pose representation and moved into the downstream training and publishing workflow.

4 The SignVerse-2M Corpus

4.1 Corpus Characteristics and Research Value

Data Organization and Corpus Value. SignVerse-2M preserves the complexity of open-world data. These videos come from different cameras, framing conditions, subtitle habits, and content genres, so the input conditions are inherently heterogeneous. As a result, the corpus construction process must prioritize stable throughput, error recovery, and traceable runtime states on heterogeneous inputs rather than assuming a clean and controlled setting. For downstream modeling, the resulting corpus is also closer to the data distribution that real deployments are likely to encounter. At the organizational level, SignVerse-2M adopts a per-video directory structure. Each ‘video_id‘ corresponds to an individual directory containing subtitle files, a structured caption JSON, the pose result ‘poses.npz‘, and a ‘.complete‘ marker indicating successful processing. At the global level, the system also maintains ‘stats.npz‘ as a lightweight state index that records the processing status of each sample across metadata, subtitle, download, process, and upload stages. This design enables both local recomputation at the level of individual videos and corpus-scale progress tracking and failure recovery. The current manifest covers roughly 39K videos and about 2M segments. At this scale, the primary determinant of corpus usability is not a highly detailed description of individual samples, but whether the processing system can reliably complete acquisition, cleaning, pose extraction, serialization, and archiving. Accordingly, SignVerse-2M emphasizes long-running execution, batch processing, status tracking, and incremental updates rather than collapsing all logic into a one-shot offline preprocessing script. From the repository implementation, downloading, pose extraction, and uploading are all

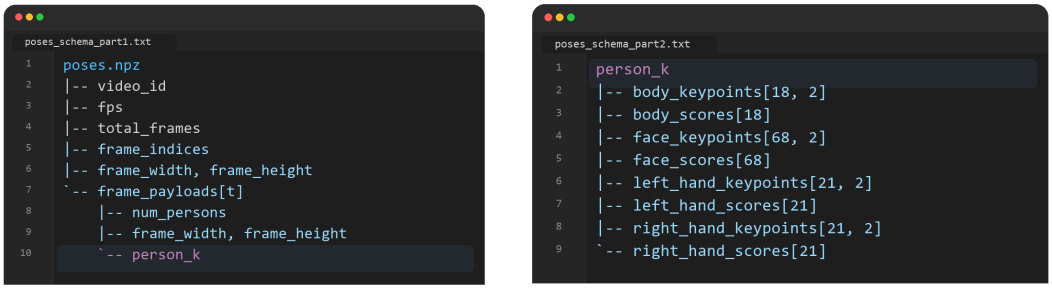

Figure 4: **Schema of the released 'poses.npz' payload.** The stored format is person-centric: each frame payload records the number of detected signers and organizes body, face, left-hand, and right-hand keypoints together with their confidence scores under each 'person_k'. This structure is the native representation released by SignVerse-2M and is the basis from which visualization scripts derive OpenPose-style aggregated renderings.

separated into distinct stages and can be executed on selected subsets of samples, which leaves room for future expansion and maintenance.

Cross-Lingual Scale. Figure 3 gives a direct view of the language distribution inherited from YouTube-SL-25. The logarithmic y-axis makes the long-tail structure visible: a small number of languages contribute a large fraction of the total hours, while many other languages remain present at much smaller scales. For SignVerse-2M, this distribution is not a side detail but a core property of the corpus, because it shapes multilingual training, transfer behavior, and the practical difficulty of benchmarking across languages.

4.2 DWPose and Visualization

DWPose Representation and Visualization. SignVerse-2M uses DWPose as the backend for pose extraction. For each frame, the system extracts body, hand, and facial keypoints together with their confidence scores. In the aggregated format, 'poses.npz' stores not only video-level metadata but also per-frame payloads; each payload records the number of people, frame dimensions, and person-wise body, face, left-hand, and right-hand keypoints and scores. This representation allows downstream methods to consume temporal inputs directly at the keypoint level without re-decoding the original videos. The dataset repository also provides visualization scripts that reconstruct 'poses.npz' into inspectable skeletal renderings. Specifically, the scripts can read either aggregated or per-frame NPZ outputs and generate multiple visualization styles, including control-style skeletons, OpenPose-style renderings, and previews overlaid on the original video. The underlying drawing functions render body, hand, and face keypoints separately while using confidence scores to modulate the display. As shown in Figure 4, the released 'poses.npz' format stores pose data in a person-centric structure rather than as a single flat OpenPose-style array. These visualization interfaces are not intended as part of the training input itself; instead, they serve as inspection and debugging tools for quickly identifying missing keypoints, multi-person interference, local jitter, and abnormal frames.

Processing Skeleton. From an implementation perspective, the more informative summary is the structure of the stored pose payload rather than another pipeline sketch. The released 'poses.npz' format can be summarized by the following schema-like pseudocode:

This organization makes explicit that SignVerse-2M stores pose data in a person-centric DWPose-style payload, with body, face, and both hands grouped under each detected signer at each frame. The accompanying visualization scripts can reorganize the same information into an OpenPose-style aggregated representation for rendering, but the released corpus itself keeps the per-person structure because it is easier to inspect, serialize, and reuse in downstream modeling.

Model	Language	DEV SET					TEST SET				
		BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
SignDW Transformer (40M)	ASL	11.27	17.11	25.82	34.97	35.05	9.65	15.51	24.40	33.85	33.71
SignDW Transformer (1.2B)	ASL	11.72	17.56	26.12	35.82	35.89	10.10	15.96	24.70	34.30	34.26
SignDW Transformer (40M)	BSL	9.58	15.32	24.05	33.80	33.42	8.35	14.04	22.72	32.58	32.08
SignDW Transformer (1.2B)	BSL	10.06	16.06	25.07	35.02	34.77	8.80	14.49	23.12	33.23	33.03
SignDW Transformer (40M)	GSL	9.82	15.68	24.48	34.23	33.98	9.69	15.48	24.29	33.82	33.58
SignDW Transformer (1.2B)	GSL	10.70	16.80	25.94	35.73	35.55	10.14	16.19	25.28	34.97	35.04
SignDW Transformer (40M)	DSGS	10.16	16.07	25.02	34.73	34.25	9.12	14.96	23.77	33.10	33.09
SignDW Transformer (1.2B)	DSGS	10.70	16.70	25.61	35.49	35.56	9.57	15.41	24.07	33.73	33.87

Table 2: **Text-to-pose baseline results on four sign languages.** We report results for **SignDW Transformer (40M)** and **SignDW Transformer (1.2B)** on ASL, BSL, GSL, and DSGS.

5 Evaluation & Baseline

Given the constraints of space and computational resources, we focus in this paper on a sign language generation baseline, as this setting more directly highlights the strengths of our dataset.

5.1 Setup

Task. Sign language generation aims to synthesize sign language motion from complex textual inputs. In this paper, our quantitative evaluation is carried out in pose space: the generated DWPose sequence is translated back into spoken-language text, and the recovered sentence is then compared with the original input. This back-translation protocol provides a broad proxy for semantic fidelity while avoiding additional variation introduced by downstream video rendering Saunders et al. [2020b].

We evaluate the pipeline at two levels. First, we measure the performance of the text-to-DWPose model itself using automatic back-translation metrics. Second, we feed the generated pose videos into modern pose-driven rendering models to synthesize final videos and inspect the overall generation quality qualitatively. The rendered videos are therefore used as a compatibility and visualization check rather than as part of the automatic evaluation protocol.

For pose-space back-translation, we use SignX Fang et al. [2026], a recent sign language translation model trained on pose data in the corresponding language. Depending on the dataset, its BLEU-4 performance typically falls in the range of 25–28, making it a strong translation model for this evaluation setting.

Metrics. `(i) BLEU-n score` measures the similarity between the generated translation and the reference text based on n-gram overlap. Higher scores indicate that the prediction is closer to the reference sentence. Larger values of n impose stricter requirements on local fluency and phrase-level consistency Papineni et al. [2002].

`(ii) ROUGE score` Lin [2004] is similar in spirit to BLEU, but places greater emphasis on coverage and overlap with the reference text. A higher ROUGE score indicates that the generated text is more consistent with the reference and is therefore more complete and accurate.

5.2 Results.

Table 2 summarizes the text-to-pose results on four representative sign languages: ASL, BSL, GSL, and DSGS. For all runs, we train the baseline for 200K steps with batch size 32 and an initial learning rate of 0.001 under the same multilingual DWPose interface, so that differences across rows mainly reflect model scale and language-specific data conditions rather than changes in the optimization setting. Overall, the proposed baseline achieves stable performance across all four settings, which suggests that the unified DWPose representation is sufficient to support multilingual text-to-pose modeling under a shared training and evaluation interface. Across languages, the larger **SignDW Transformer (1.2B)** model is consistently stronger than the **SignDW Transformer (40M)** model, with the clearest gains appearing in BLEU-4 and BLEU-3, indicating better preservation of longer local motion patterns after back-translation. At the same time, the gap between languages remains non-negligible: ASL and GSL are relatively stronger, while BSL and DSGS remain more challenging, especially on the test split. We view this pattern as consistent with the multilingual and open-world nature of SignVerse-2M. The benchmark is therefore not only a proof of feasibility, but also a first

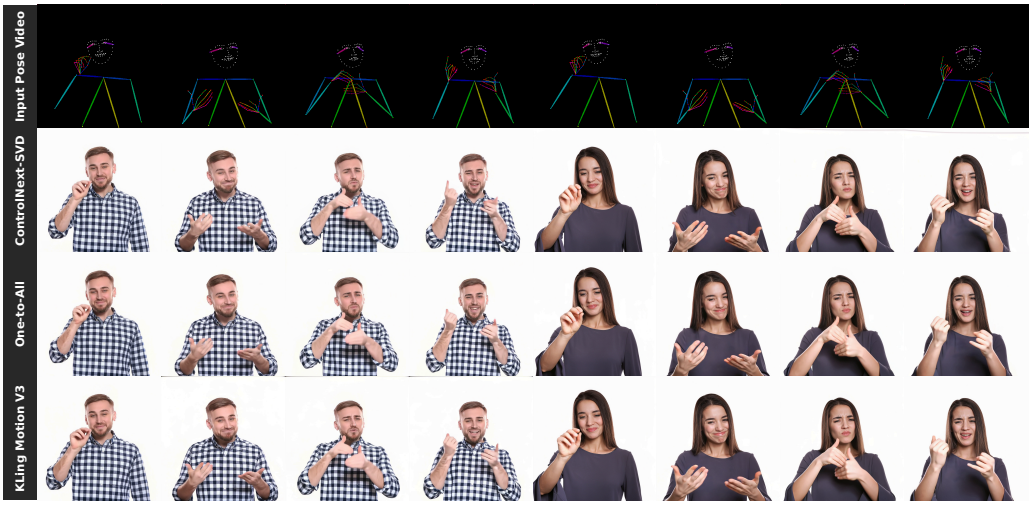

Figure 5: **Qualitative comparison of pose-conditioned sign video rendering.** The first row shows the input DWPose video. To remain compatible with the model-specific preprocessing pipelines of different renderers, we prepend a body-shape mask before feeding the pose video into each method. The next three rows compare **ControlNext-SVD** Peng et al. [2025], **One-to-All** Shi et al. [2025], and **KLing Motion V3** Team et al. [2025] under the same pose sequence. For each rendering row, the left half shows the male signer and the right half shows the female signer.

indication that model scale and language-specific data conditions both matter under a pose-native sign generation setting.

Figure 5 provides a complementary qualitative view of this result. Beyond back-translation scores, the examples show that the DWPose-based representation used in SignVerse-2M can be interfaced directly with several modern pose-conditioned rendering systems, including both commercial and open-source pipelines, while still preserving the signer's coarse motion structure and temporal progression. This point is important for our dataset motivation: earlier sign-language pipelines often relied on representations such as MediaPipe pose, OpenPose-style subsets, or SMPL-derived intermediate formats, which are useful for analysis but are not naturally aligned with the control interfaces adopted by recent image-to-video and talking-human generation models. In contrast, the representation used here makes the downstream rendering step substantially easier to connect, compare, and reuse, which is precisely the kind of systems-level compatibility that a pose-native benchmark should provide.

6 Conclusion

We present SignVerse-2M, a large-scale multilingual pose-native dataset for sign language research. Unlike existing sign language resources that are primarily derived from laboratory settings, the core contribution of SignVerse-2M is neither simply to scale up the amount of video data nor to introduce a new pose estimation algorithm. Instead, it systematically converts publicly available multilingual sign language videos from open environments into a unified DWPose representation, and builds around this representation the data interface and evaluation foundation needed for sign language tasks. Through this resource, we aim to shift the focus of sign language generation research away from heterogeneous RGB video preprocessing and incomparable experimental settings toward motion representation, cross-lingual transfer, and evaluation itself.

Overall, we hope that SignVerse-2M will provide not merely a collection of samples, but a new research interface. This resource helps move the community's focus from performance validation in idealized laboratory settings toward robust modeling and deployment evaluation under real-world conditions, thereby offering a more unified, transparent, and comparable foundation for future sign language research.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up Co-articulated Sign Language Recognition using Mouthing Cues. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset. 2021.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text, 2024. URL https://arxiv.org/abs/2312.02702.

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*, pages 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 Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. doi: 10.1109/CVPR.2018.00812.

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.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now, 2019. URL https://arxiv.org/abs/1808.07371.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

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.

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)*, 2021a.

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)*, 2021b.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. Signllm: Sign language production large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6622–6634, October 2025a.

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, Hongwei Yi, Carol Neidle, and Dimitris N. Metaxas. Signx: The foundation model for sign recognition, 2025c. URL https://arxiv.org/abs/2504.16315.

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

Sen Fang, Yalin Feng, Chunyu Sui, Hongbin Zhong, Hongwei Yi, and Dimitris N. Metaxas. Signx: Continuous sign recognition in compact pose-rich latent space, 2026. URL https://arxiv.org/abs/2504.16315.

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.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, 2014.

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*, 2023.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

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, 2019.

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

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*, pages 1459–1469, 2020.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, July 2004.

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.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos, 2024. URL https://arxiv.org/abs/2304.01186.

Amit Moryossef and Mathias Müller. Sign language datasets. https://github.com/sign-language-processing/datasets, 2021.

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

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*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation, 2025. URL https://arxiv.org/abs/2408.06070.

Manny Rayner, Pierrette Bouillon, Sarah Ebling, Johanna Gerlach, Irene Strasly, and Nikos Tsourakis. An open web platform for rule-based speech-to-sign translation. In *54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 162–168, 2016.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video, 2020a. URL https://arxiv.org/abs/2011.09846.

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)*, 2021.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Shijun Shi, Jing Xu, Zhihang Li, Chunli Peng, Xiaoda Yang, Lijing Lu, Kai Hu, and Jiangning Zhang. One-to-all animation: Alignment-free character animation and image pose transfer, 2025. URL https://arxiv.org/abs/2511.22940.

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.

Garrett Tanzer, Biao Zhang, David Uthus, and Manfred Georg. YouTube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus, 2024.

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.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-omni technical report, 2025. URL https://arxiv.org/abs/2512.16776.

David Uthus, Garrett Tanzer, and Manfred Georg. YouTube-ASL: A large-scale, open-domain american sign language-english parallel corpus, 2023.

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*, pages 243–246, 2010.

Harry Walsh, Ed Fish, Ozge Mercanoglu Sincan, Mohamed Ilyes Lakhal, Richard Bowden, Neil Fox, Bencie Woll, Kepeng Wu, Zecheng Li, Weichao Zhao, Haodong Wang, Wengang Zhou, Houqiang Li, Shengeng Tang, Jiayi He, Xu Wang, Ruobei Zhang, Yaxiong Wang, Lechao Cheng, Meryem Tasyurek, Tugce Kiziltepe, and Hacer Yalim Keles. Slrtp2025 sign language production challenge: Methodology, results, and future work, 2025. URL https://arxiv.org/abs/2508.06951.

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*, pages 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*, pages 5109–5119, 2022.

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text, 2024. URL https://arxiv.org/abs/2406.07119.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation, 2023. URL https://arxiv.org/abs/2305.13077.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance, 2025. URL https://arxiv.org/abs/2406.19680.

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)*, pages 1316–1325, 2021.

Inge Zwitserlood, Margriet Verlinden, Johan Ros, and Sanny Van Der Schoot. Synthetic Signing for the Deaf: Esign. In *Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment (CVHI)*, 2004.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explicitly state the paper's scope as a multilingual pose-native dataset, its task interface, and a simple baseline; see the abstract and Sections 1 and 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations of the pose-native representation, including fine-grained hand information, non-manual features, and linguistic completeness; see the abstract, Section 1, and the concluding discussion in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the data construction pipeline, the stored pose format, the evaluation setup, and the baseline task interface in enough detail to support the main feasibility claims; see Sections 3, 4, and 5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The submission is accompanied by dataset and code resources for the released benchmark and processing pipeline, and these materials are intended to provide the practical entry points needed to reproduce the main experiments and inspect the released asset.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the task definition, evaluation protocol, data splits, and core baseline settings, including the shared multilingual setup, training schedule, batch size, and learning rate used for the reported results; see Section 5 and Table 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The current experiments report point estimates only; no error bars, confidence intervals, or significance tests are included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not currently report detailed hardware, memory, runtime, or total compute usage for the baseline experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the work is consistent with the NeurIPS Code of Ethics, and the paper discusses scope, deployment relevance, and limitations of the released resource.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: The paper motivates accessibility and real-world robustness, but it does not yet provide a dedicated broader-impact discussion covering both positive and negative societal effects in a systematic way.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper does not yet include a dedicated safeguards section describing release controls, risk mitigation, or filtering procedures for potentially sensitive scraped web data.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing assets are cited in the paper, and the release terms for key reused resources are identified in the accompanying materials. In particular, the Google Research repository hosting YouTube-SL-25 states that datasets are released under CC BY 4.0 and source files under Apache 2.0, while the underlying videos remain subject to the original YouTube content licenses and platform terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper documents the new pose-native dataset interface, processing pipeline, pose storage schema, and benchmark setting in Sections 3–5, and the released project materials are intended to accompany the asset. The released codebase is distributed under the MIT license, while our released dataset annotations and metadata are distributed under CC BY-NC 4.0 with accompanying usage documentation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not report crowdsourcing studies or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not report research with human subjects requiring IRB-style review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core contribution of the paper is the dataset, processing interface, and baseline evaluation setup rather than an LLM-based method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.