

Diffusion Model for Chilean Sign Language

Joaquín Figueroa

*Departamento de Ciencias de la Computación
Universidad de Chile*

Santiago, Chile

joaquin.figueroa.m@ug.uchile.cl

Miguel Yáñez

*Escuela de Ingeniería Informática
MEDING Interdisciplinary Center*

Universidad de Valparaíso

Valparaíso, Chile

miguel.yanez@postgrado.uv.cl

Ivan Sipiran

*Departamento de Ciencias de la Computación
Universidad de Chile*

Santiago, Chile

isipiran@dcc.uchile.cl

Ana Aguilera

*Escuela de Ingeniería Informática
MEDING Interdisciplinary Center*

Universidad de Valparaíso

Valparaíso, Chile

ana.aguilera@uv.cl

Abstract—Communication barriers limit the access of deaf communities to information and services, highlighting the need for technologies that support sign language production. In Chile, this challenge is compounded by the scarce availability of Chilean Sign Language (LSCh) corpora, which hinders training and evaluation of generative models. In this work, we implement the first stage of a text-to-video LSCh translation pipeline: a conditional signer video generator inspired by SignViP, which synthesizes video sequences from a reference image and structured pose conditions. We introduce COTRA-LSCh, a new LSCh corpus under construction, and design a processing pipeline that extracts body and facial poses with DWPose and 3D hand representations with HaMeR, standardizing both modalities into visual conditioning signals. The generator is trained in two stages, first learning identity and condition alignment on short clips and then incorporating temporal coherence via motion modules. We evaluate the resulting system through qualitative inspection and standard video metrics, and we compare a model trained on COTRA-LSCh against an analogous model trained on RWTH-T. Results indicate that COTRA-LSCh generations preserve global appearance and body motion consistency, achieving stronger temporal and structural agreement at the sequence level, with lower FVD (1389.03 vs. 1526.58) and slightly higher SSIM (0.66 vs. 0.64). Across both qualitative and quantitative evidence, the main artifacts concentrate on hands under contact/occlusion and limited facial expressiveness. Overall, these findings suggest that strong structural conditioning can anchor LSCh motion and appearance at the sequence level, while improving fine-grained hands and facial dynamics remains the main direction for future work.

Index Terms—Sign Language Production, Chilean Sign Language, Stable Diffusion, AnimateDiff Motion Model

I. INTRODUCTION

Communication is an essential pillar of life in society. However, for people with hearing disabilities, this fundamental right is constantly violated by the lack of accessibility in everyday channels of interaction [1], [2]. Communication barriers prevent many deaf people from fully accessing education, the world of work, health services, and, in general, the various areas of social participation [2], [3]. As a result, much of the cultural and social richness that makes up our community life

is inaccessible to them, leading to exclusion and reinforcing historical inequalities.

It is important to recognize that sign language is not merely a visual translation of Spanish, but a complete linguistic system with its own grammar, syntax, and semantics [4], [5]. A signer may not understand written or spoken Spanish, just as a Spanish speaker without training may not understand sign language. This linguistic mismatch deepens the distance between communities and raises the urgency for technological solutions to bridge the gap.

In this context, the idea of building an automatic text-to-video translator in LSCh, known as sign language production, arose. The motivation behind this proposal is to offer a tool that can generate realistic and expressive videos in which a signing avatar conveys information written in Spanish in its equivalent in LSCh. If successful, such a system would substantially increase accessibility to information in areas as diverse as education, health, public services, and digital content consumption.

Currently, the availability of corpora for LSCh remains very limited, which hinders the systematic training of deep learning models for machine translation [6]. In light of this gap, a team of five linguists is developing the first comprehensive LSCh corpus, designed to be as general and representative as possible. This dataset seeks to cover a wide variety of subject areas and signs, and is characterized by high-quality audiovisual recordings with a resolution of 1980x1080 and a rate of 120 FPS. It is estimated that the dataset will consist of a total of 12,000 clips, surpassing in size widely used corpora such as PHOENIX [7], and constituting a sufficient basis for training large-scale deep learning models. It is hoped that this resource will become a fundamental pillar for experimentation and development of different approaches to Sign Language Production (SLP) applied to LSCh [8].

The technical challenge underlying this idea is complex. The process does not simply consist of “converting” words into signs, but rather translating the text into an interme-

mediate representation that captures the semantic content of the sentence. This representation, rather than being directly visible, functions as a conceptual bridge that is then delivered to a generator model capable of producing a coherent and expressive video. In other words, the system must understand the meaning of the text and, based on that, create an animation that reproduces not only hand movements, but also body posture, facial expressions, and the temporal fluidity of the gesture. These elements are fundamental, since in sign language, expressions and bodily prosody provide key semantic information [9]. A translator who produces rigid, artificial, or poorly detailed videos runs the risk of conveying erroneous or incomplete meanings, creating confusion rather than accessibility.

To address this challenge, this paper proposes using a state-of-the-art methodology inspired by a recent development by Chinese researchers called SignViP [10]. Adapting this technique to the Chilean context, applying it to LSCh, not only represents an academic and technical challenge, but also a valuable social contribution to inclusion and equity. This paper presents the results of the implementation of the first stage of the system: a generator model capable of synthesizing realistic video sequences of a signer based on a reference image and a time series of body poses.

This paper is structured as follows. Section 2 presents related work in the context of Sign Language Production. Section 3 develops the proposed architecture followed by its validation in Section 4. Comparison and analysis are presented in Section 5. Finally, Section 6 presents conclusion and future work.

II. RELATED WORK

The automatic generation of sign language videos, known in the literature as Sign Language Video Generation (SLVG), is an emerging field of research that seeks to combine advances in artificial intelligence, natural language processing, and computer vision. In recent years, various approaches have been proposed to address this challenge, which have marked important milestones thanks to deep learning, but have also revealed limitations that open the door to new solutions [11], [12].

Figure 1 illustrates a high-level, black-box view of the SLVG problem.

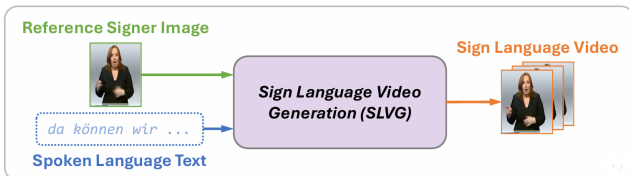


Fig. 1. Black-box formulation of sign language video generation (SLVG). Cropped from Fig. 1 in [10].

The present work covers the generation of videos of signing avatars with identity preservation based on a temporal sequence of poses and an image of a person.

Initial methods involved generating videos from a single intermediate condition, usually sequences of skeletons representing the basic body posture of the signer [13]–[15]. The idea was to translate the text into this skeletal representation and then generate the video from it. Although this approach allowed for a first step toward the automatic production of signs, it soon became clear that the generated videos lacked naturalness and expressiveness [13], [16]. This was because the skeletons were too simple: they did not capture crucial details such as precise hand movements or facial expressiveness, both of which are fundamental in sign language for differentiating meanings [11], [17]–[19]. The result was videos that were not very fluid and had low communicative value for the deaf community.

The single-intermediate-condition pipeline described above is summarized in Fig. 2.

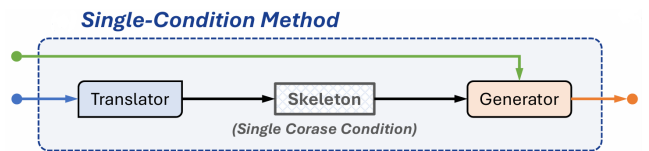


Fig. 2. Multi-condition method for sign language video generation. Cropped from Fig. 1 in [10].

The most representative example of the intermediate condition generation modality is SignGan [14] (2020), a foundational model that set the standard for photorealistic generation. The architecture is not completely end-to-end in a single step. First, it uses a Transformer-based network combined with a Mixture of Density Networks (MDN) to translate text into a sequence of skeletons (poses). The second stage uses a conditional GAN (Pix2PixHD or Video-to-Video type) that takes the sequence of skeletons and a reference image (style) to generate the final video. Its key architectural innovation was not the generative engine itself, but the introduction of keypoint-based loss. This loss function operates in the space of the keypoints of the hand to penalize motion blur, forcing the generator to produce sharp hands, something that traditional GANs (Generative Adversarial Networks) failed to do [14].

Given these shortcomings, a second approach emerged: direct multi-condition methods [20]. In this case, instead of working with a single skeleton, the system attempted to simultaneously generate more detailed conditions, such as fine body poses and three-dimensional models of the hands. In theory, this increase in information would allow for the production of much more realistic and expressive videos. However, in practice, this method faced a critical obstacle: the high dimensionality and complexity of the conditions generated noisy and error-prone translations [10], [20]. When attempting to move directly from text to these multiple detailed representations, the system failed to maintain consistency and some of the meaning was lost during translation.

An overview of direct multi-condition methods is shown in Fig. 3.

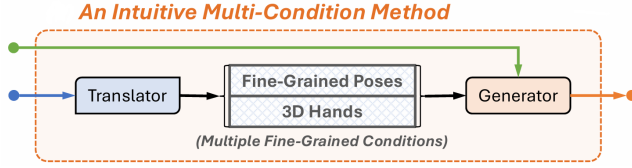


Fig. 3. Multi-condition method for sign language video generation. Cropped from Fig. 1 in [10].

SignGen [20] (2024) marks the shift to Latent Diffusion Models (Video LDM), overcoming the limitations of GANs through an iterative noise removal process. It uses a modified U-Net architecture within a latent space. Unlike SignGAN, it does not generate pixels directly in one go, but rather “denoises” (cleans noise) iteratively guided by conditions.

Its major architectural difference is the motion condition fusion module. It not only uses the skeleton (pose) like SignGAN, but also injects three simultaneous continuous signals: Optical Flow (for motion), Depth (for 3D structure), and Pose. It incorporates a Spatio-Temporal-Semantic Attention mechanism within the U-Net to ensure that the generated video is not only temporally smooth, but that the meaning of the text aligns with the gestures in each frame [20].

Faced with these limitations, the most recent proposal, SignViP [10], introduces a key innovation: instead of translating text directly into complex conditions, it uses a more compact and robust intermediate representation based on discrete tokens. A multi-condition token translator is introduced that translates spoken text into an intermediate representation that explicitly includes fine-grained poses and 3D hands, before passing it to the diffusion model.

The overall SignViP approach is depicted in Fig. 4.

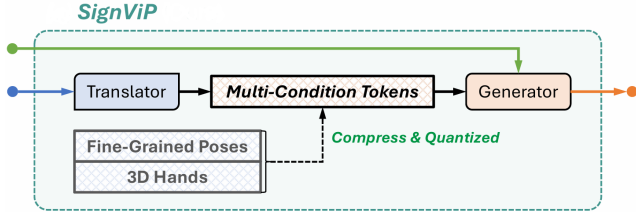


Fig. 4. SignViP method for sign language video generation. Cropped from Fig. 1 in [10].

This abstraction step allows the semantic richness of the message to be preserved without being affected by the noise of high-dimensionality conditions. These tokens then serve as a guide for a video diffusion model, which is responsible for generating realistic and expressive animations. This approach has been shown to outperform previous methods [10], achieving results with greater temporal consistency, better visual quality, and greater semantic fidelity to the original message.

Table I and Table II summarize the visual quality results reported in Table 3 for RWTH-2014T and How2Sign [7], [10], [16].

TABLE I
VISUAL QUALITY SUMMARY ON RWTH-2014T (FROM TABLE 3 IN [10]).
LOWER IS BETTER FOR FID/FVD, HIGHER IS BETTER FOR IDS.

Method	FID ↓	FVD ↓	IDS ↑
SignGAN	547.90	1431.38	0.463
SignGen	644.06	1715.32	0.515
SignViP	508.91	1025.45	0.571

TABLE II
VISUAL QUALITY SUMMARY ON HOW2SIGN (FROM TABLE 3 IN [10]).
LOWER IS BETTER FOR FID/FVD, HIGHER IS BETTER FOR IDS.

Method	FID ↓	FVD ↓	IDS ↑
SignGAN	667.44	2766.97	0.5377
SignGen	815.69	3538.49	0.5390
SignViP	575.67	2207.67	0.624

III. MATERIALS AND METHOD

This section describes the materials and methodology employed in this work.

A. Materials

1) *Corpus Compilation and Annotation*: In parallel with the methodological development of this project, a new corpus of LSCh is being compiled under the direction of Ana Aguilera as part of the Fondecyt Regular 1250344 project “A Translation System Based on a Hybrid Neural Machine Applied to Sign Language Production.” This effort involves recording sessions with 30 deaf individuals, each lasting approximately one hour, with the aim of capturing inter-subject variability and a visual signal rich enough for training generative models. Subsequently, the material undergoes an annotation process designed to mark glosses and direct translations, carried out by a team of five linguists, including interpreters and deaf individuals, which allows for a balance between linguistic criteria and situated knowledge of the community. Currently, the recordings are obtained as raw video in 1920×1080 format at 120 FPS, and is expected to have an initial dataset of approximately 12,000 clips (estimated at 400 segments per participant) by the end of May, providing a quantitative basis for training and evaluating the proposed modules. We will refer to this dataset as COTRA-LSCh throughout the paper. Representative sample frames are shown in Fig. 5.

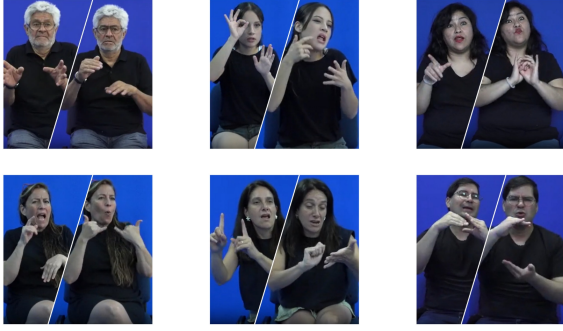


Fig. 5. Sample frames from the Chilean Sign Language corpus under construction.

2) *Computational Resources*: The experiments were conducted using resources from the Department of Computer Science at the Faculty of Physical and Mathematical Sciences of the University of Chile, provided as a shared computing system. This infrastructure consists of multiple nodes with heterogeneous capabilities and a common centralized storage system, which allows the dataset and source code to be stored in a single location accessible from any server. Consequently, preprocessing, training, and inference can be distributed across available machines without manual data replication, facilitating operations on a large corpus.

Table III summarizes the main specifications of the computational resources used:

TABLE III
TECHNICAL SPECIFICATIONS OF THE RELELA MACHINES USED IN THIS WORK.

RAM	Processor	GPUs	Objective
128 GB DDR4	AMD EPYC 7282	3 × NVIDIA RTX A5000 (24 GB)	Preprocessing
128 GB DDR4	AMD Ryzen 7 5700X	2 × NVIDIA RTX A6000 (48 GB)	Training
128 GB DDR5	Intel Core i7-12700K	2 × NVIDIA GeForce RTX 3090 (24 GB)	Inference

3) *Pre-trained Models*: In this implementation, pre-trained models are reused as functional building blocks, with the aim of reducing training costs and leveraging established visual representations. Specifically:

- **Stable Diffusion v1.5**: It is used as a generative backbone in latent space and serves as the starting point for fine-tuning for the sign language domain [21], [22]. Having been trained on large collections of images, it provides robust priors regarding human appearance, lighting, and textures, which facilitates photorealistic synthesis [21], [22]. At the architectural level, it relies on a U-Net for iterative denoising, a scheduler that controls noise dynamics, and a text encoder/tokenizer for conditioning [21].
- **Variational Autoencoder**: It is used as an encoding/decoding mechanism between the pixel space and the compressed latent space in which the diffusion model operates [21], [23]. During training, it transforms high-dimensional videos into smaller latent representations, thereby speeding up computation and reducing memory usage. During inference, it performs the reverse

operation, decoding the generated latent samples into observable video frames.

- **AnimateDiff Motion Module**: It is integrated as a motion module that can be coupled with U-Net to introduce explicit temporal coherence between frames [24]. The rationale behind this is that standard image diffusion models do not natively model time [24]; therefore, this component learns spatiotemporal dependencies that stabilize the dynamics of the video. In practice, it transforms a predominantly T2I generator into one capable of producing smoother and more consistent T2V sequences.
- **DWPose**: It is used in the knowledge extraction stage to obtain consistent 2D skeletons from raw video data [25]. This extractor outputs keypoints for the body, hands, and face, providing a structured signal that serves as a strong prior for generation (for example, in ControlNet-style architectures [26]). In this way, the generative model can focus on appearance and identity, while the kinematics are anchored by the estimated pose.
- **HaMeR (Hand Mesh Recovery)**: It is used to reconstruct a dense 3D representation of the hand from 2D video [27], complementing extractors based solely on keypoints. By estimating the complete mesh and kinematic parameters (orientation of the wrist and phalangeal joints), it provides a more informative description of the hand configuration. This is particularly relevant in sign language, where subtle variations in hand shape can alter meaning and therefore require higher-fidelity representations [11], [17].

B. Proposed Method

Figure 6 provides a high-level overview of the proposed pipeline, from raw corpus preparation and condition extraction to the two-stage training strategy and final conditional video generation. This diagram serves as a roadmap for the following subsections, where each module is described in detail.

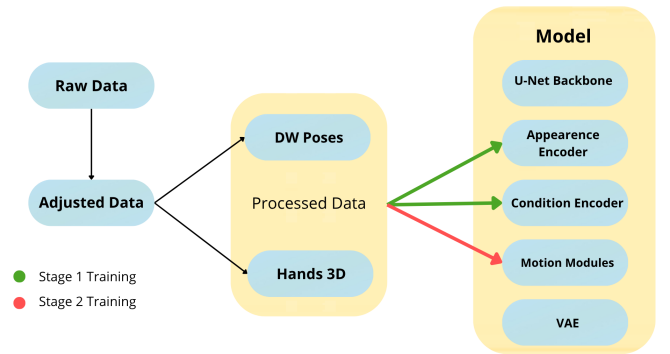


Fig. 6. Overview of the proposed method, where are seen the stages of data and which parts of the model are trained during each process of the training.

1) *Preprocessing*: The preprocessing of the material was designed to enable training with limited computational resources, while preserving the visual information relevant to

sign language comprehension. First, since the original recordings are in FHD resolution, operating directly at that scale would entail a substantial increase in computational and memory costs, in addition to requiring architectures with greater capacity to adequately process the signal; for this reason, the videos were rescaled to 640×480, a more manageable resolution that, at the same time, remains higher than that used in RWTH-T [7]. Second, considering that the labeling process is not yet complete, we chose to use the raw videos to build the training set, partitioning each approximately one-hour session into random clips ranging from 4 to 10 seconds; as part of the descriptive analysis, we examine the resulting distribution of durations across the entire dataset. Finally, once the clips were generated, the dataset was divided into three partitions (train, dev, and test) in an 85ratio, yielding 7,708, 906, and 455 videos per split, respectively.

2) *Processing*: The processing stage transforms raw audio-visual data into structured representations that can be used as training data. First, DWPose and HaMeR are applied to extract, respectively, high-resolution body and facial poses, along with a 3D representation of the hands; due to the large size of the dataset, this extraction is performed in parallel across multiple GPUs to reduce the total computation time. Second, the format of the conditions is standardized so that the diffusion model can learn from visual signals: while DWPose provides keypoints that can be rendered as images using `draw_pose` during training, HaMeR produces `.pkl` files with hand parameters that cannot be directly consumed by the diffusor, thus they are first converted to a video-like format. Since this conversion relies solely on the CPU and is particularly computationally expensive, parallelization is again employed to accelerate pre-training processing. An overview of the resulting visual conditioning signals is shown in Fig. 7.

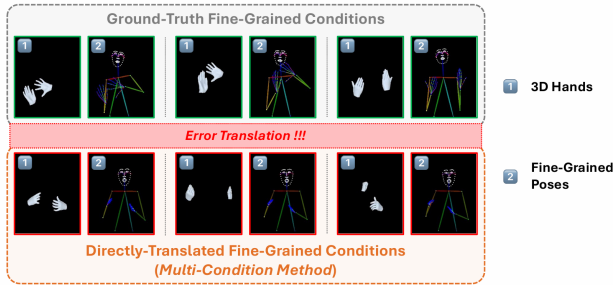


Fig. 7. Illustration of the fine-grained pose and 3D Hands conditions used by the pipeline. Cropped from Fig. 1 in [10].

3) *Training*: In the first training stage, the goal is for the system to learn to reconstruct plausible and consistent frames based on two complementary sources of information: a reference image that establishes the signer’s appearance and identity, and pose-based conditions (skeleton/face plus hands) that describe the target body configuration. To achieve this, training is conducted using very short clips (4 frames), deliberately avoiding the requirement for long-term temporal coherence and focusing on stabilizing the relationship between

conditions and appearance. In practical terms, the dataset provides for each sample the target RGB video along with its aligned visual conditions, as well as a reference frame taken outside the target segment to reduce the risk of the model directly “copying” the response. During this stage, the generative backbone (VAE and U-Net) is frozen, and the encoders that integrate identity and conditions are optimized, selecting the best checkpoint via validation on the “dev” split and selecting the best model based on maximizing the SSIM metric [28].

The second stage explicitly introduces motion learning, extending the training from the generation of isolated frames to longer, temporally coherent sequences. Instead of re-training the entire model, we start from the weights obtained in Stage 1 (in particular, the encoders that have already learned to preserve identity and interpret conditions) and freeze most of the architecture, training only the temporal modules (motion modules). This approach reduces the risk of degrading previously learned visual quality and focuses the adjustment capacity on the dynamics between frames, using longer clips (24 frames). Evaluation is performed again on the development set, but now by comparing generated full-length videos against their real-world counterparts using FVD [29]; the best model is selected based on the lowest FVD, and only the parameters associated with the motion modules are retained.

4) *Inference*: During inference, the system generates a conditional video by running the SignViP pipeline in two stages. For each sample, the target clip (for evaluation) is loaded along with its associated conditions—poses estimated with DWPose and the hand sequence generated by HaMeR. To present the model with homogeneous visual signals, DWPose keypoints are rendered frame-by-frame as images using `draw_pose`, while HaMeR is used directly as an RGB video. Finally, a reference frame (the one at 1 second into each video) is set to anchor the signer’s identity, and the diffusion process is run on the sequence using the same time window as in training; in our experiments, we used 20 diffusion steps and moderate guidance (guidance scale 3.5) to balance fidelity to the conditions and naturalness of movement.

5) *Metrics*: To quantify the quality of the generated videos, an evaluation orchestrator was implemented that systematically compares two folders by matching files by name. The videos are loaded using OpenCV and normalized to a consistent format (RGB and values in the range [0,1]) to ensure comparability between metrics. Additionally, when a metric requires it, resolution compatibility is checked by filtering the pairs and working with a common target resolution, thereby avoiding biases and errors caused by inconsistent dimensions.

The evaluation includes the following metrics:

- **FVD (Fréchet Video Distance)**: FVD evaluates the overall similarity between real and generated videos by comparing distributions of spatiotemporal representations extracted using a pre-trained I3D backbone [29], [30]; in practice, it simultaneously captures both appearance and motion dynamics, where lower values indicate greater similarity to the real dataset. To make the calculation

more robust, a minimum number of frames per clip is required; the length of each pair is equalized by truncating the common sequence; and the data is processed in batches, reporting a weighted average over the evaluated set.

- **FID (Fréchet Inception Distance):** FID is calculated at the frame level to estimate average visual realism by comparing the distribution of Inception features between real and generated frames; since it does not explicitly incorporate time, it complements FVD by isolating the quality of appearance on a per-frame basis [31]. In this implementation, `clean-fid` is used [32], and frames are extracted in “streaming” mode to avoid loading all the data into memory; lower values are interpreted as better agreement with the real material.
- **SSIM (Structural Similarity Index):** SSIM measures the structural similarity between the original frame and its generated counterpart by combining information on luminance, contrast, and structure; as a result, it is useful for quantifying local fidelity without relying on strict pixel-to-pixel comparisons. It is reported as an average score across the video, with values close to 1 indicating greater preservation of the content’s structure [28].
- **LPIPS (Learned Perceptual Image Patch Similarity):** LPIPS estimates perceptual distance using activations from a deep neural network (AlexNet), capturing visual differences as a human observer would perceive them more accurately than pixel-based metrics. As a result, it can detect subtle textural degradations or artifacts even when SSIM is high; a lower value is interpreted as greater perceptual similarity [33], [34].
- **IDS (Identity Score):** To evaluate identity preservation, IDS compares facial embeddings obtained using a pre-trained recognizer (InsightFace/ArcFace) between the original and generated videos, using cosine similarity as a measure of correspondence [35]–[37]. The procedure detects faces frame by frame, aggregates a representation per video, and then compares them in pairs; if no face is detectable in either video, the pair is excluded to avoid spurious conclusions. Higher values indicate better preservation of the speaker’s identity.
- **Sharpness / NR-IQA:** Finally, we incorporate non-reference sharpness metrics (NR-IQA) to characterize common artifacts in generative video models, such as oversmoothing or loss of focus. Since these metrics do not yield a direct pairwise distance, they are reported separately for the original and generated datasets (mean and standard deviation), allowing us to detect systematic differences in sharpness attributable to the synthesis process.

IV. RESULTS AND DISCUSSION

This section discusses the results obtained by the video generator trained on COTRA-LSCh and its direct comparison with a similar model trained on RWTH-T. The analysis combines (i) qualitative inspection of the generated samples

and (ii) quantitative metrics of visual quality, identity, and temporal consistency, with the aim of identifying the strengths of the approach while also pinpointing the main areas for improvement.

A. Results

1) *Qualitative Analysis:* From a visual standpoint, the model tends to adequately preserve the signer’s overall appearance when provided with a reference image. In particular, there is good transfer of low- and mid-level attributes—such as skin tone, clothing colors, and distinctive elements (e.g., hair ties, bracelets, or patterns)—suggesting that identity conditioning successfully anchors the “macro-appearance” consistently throughout the clip. Likewise, general body kinematics are generally consistent: head movements (nodding/shaking), torso swaying, and shoulder variations are reproduced with good precision, and body posture maintains continuity between frames. An example of this transfer of appearance and posture is shown in Fig. 8.

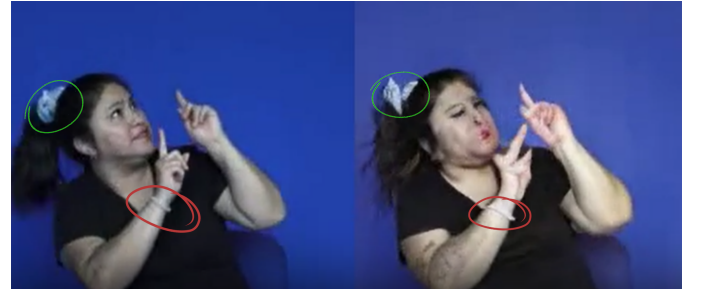


Fig. 8. Example illustrating the transfer of appearance attributes and overall body posture from the reference identity.

When it comes to hands, the system exhibits mixed results. When the hands are separated and there are no severe occlusions, the gestures tend to retain a recognizable shape and remain relatively close to the movements in the original video. However, in segments where both hands interact or come together to perform a joint gesture, visible degradations appear: the fingers may blur and the hand geometry becomes unstable, resulting in amorphous “masses” or the perception that both hands are merging. This pattern is consistent with a recurring challenge in video-driven generation: the combination of occlusions, high spatial frequency (fingers), and rapid local variations demands a level of detail that models often sacrifice in favor of global stability. Figure 9 provides qualitative examples of both well-formed hands and common failure cases under interaction and occlusion.

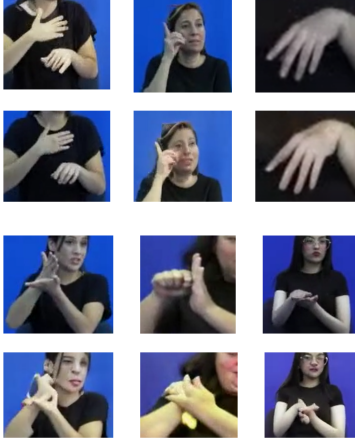


Fig. 9. Hand quality examples: well-generated hands (top) versus hand merging/blurring artifacts under interaction and occlusion (bottom).

A further relevant qualitative aspect is facial expressiveness. In many instances, the face remains relatively static, with little mouth movement and limited facial gestures, which reduces naturalness and, potentially, linguistic fidelity (given the semantic role of non-manual components) [11], [19]. In the case of COTRA-LSCh, a plausible hypothesis is that this tendency is reinforced by the construction of the training set from random clips: if a significant fraction of the segments corresponds to pauses where the person is listening to the interviewer (with a neutral expression and little articulation), the model could learn a prior that favors less expressive faces. Figure 10 illustrates this limitation through a qualitative comparison between ground-truth and generated samples.

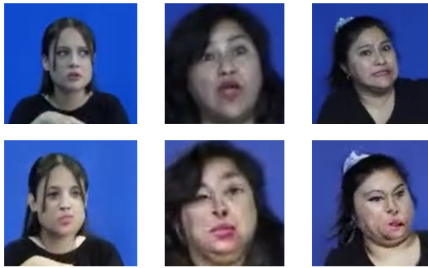


Fig. 10. Qualitative comparison of facial expressiveness between ground-truth and generated samples.

2) *Metrics*: The following are the metrics extracted from each model in their respective testing splits:

Table IV summarizes the standard video quality and identity metrics for both datasets. In addition, Table V reports reference-free sharpness metrics (NR-IQA) for COTRA-LSCh, separately on the original and generated sets.

TABLE IV
COMPARISON BETWEEN RWTH-T AND COTRA-LSCh ON STANDARD METRICS (LOWER IS BETTER FOR FVD/FID/LPIPS, HIGHER IS BETTER FOR SSIM/IDS).

Dataset	FVD ↓	FID ↓	SSIM ↑	LPIPS ↓	IDS ↑
RWTH-T	1526.57	22.94	0.64 (0.030)	0.27 (0.039)	0.57 (0.085)
COTRA-LSCh	1389.03	50.14	0.66 (0.025)	0.31 (0.030)	0.47 (0.098)

TABLE V
REFERENCE-FREE SHARPNESS METRICS (NR-IQA) ON COTRA-LSCh, REPORTED SEPARATELY FOR THE ORIGINAL AND GENERATED SETS AS MEAN (STD).

Metric	Original	Generated	Ratio (Gen/Orig)
Tenengrad	4664.28 (1416.58)	7383.55 (2253.13)	1.58
Laplacian Variance	300.87 (126.69)	282.23 (114.64)	0.94
Brenner	3032084.22 (1280421.95)	2476756.47 (1352951.05)	0.82
SAD	103650.63 (21071.51)	120440.55 (31039.22)	1.16

B. Analysis

The analysis of metrics reinforces the previous visual diagnosis and allows us to break down performance into complementary components. First, COTRA-LSCh achieves a lower FVD than RWTH-T (1389.03 vs. 1526.57; lower is better), suggesting that, at the full-sequence level, the model successfully captures plausible temporal dynamics in the LSCh domain, preserving motion coherence and global continuity. Since FVD summarizes differences in spatiotemporal representations (rather than comparing isolated frames) [29], [30], this result is consistent with the observation of stable body movements (head, torso, and posture) throughout the clip.

In contrast, when examining metrics dominated by appearance at the frame level, RWTH-T shows a marked advantage. The FID on COTRA-LSCh is significantly higher (50.14 vs. 22.94; lower is better), indicating that, frame by frame, the visual distribution of the generated images deviates more from that of the real ones. This gap is consistent with the local artifacts observed in hands during contact/occlusion and with subtle degradations that affect visual “naturalness,” even when the global structure is well preserved. In practical terms, this difference should not be interpreted solely as “lower quality,” but also as a reflection of the greater visual diversity and less controlled capture conditions in COTRA-LSCh (inter-subject variation, lighting, angles, and context), compared to the more homogeneous environment of RWTH-T.

The combined analysis of SSIM and LPIPS is particularly informative. On the one hand, COTRA-LSCh slightly outperforms RWTH-T in SSIM (0.66 vs. 0.64; higher is better), indicating good preservation of image structure (edges, luminance, and contrast) between the original and the generated image. This supports the qualitative evidence that the general body geometry, composition, and colors are consistently maintained. On the other hand, COTRA-LSCh achieves a higher LPIPS score (0.31 vs. 0.27; lower is better), which points to more noticeable perceptual differences in fine details that SSIM does not capture well. Taken together, this pattern—relatively high SSIM with unfavorable LPIPS—suggests a behavior where the “macro-structure” (posture, scene, colorimetry) is preserved

but elements of perceptual “micro-structure” (fingers, fine textures, mouth/expressions) fail, in line with visual observations.

The identity metric (IDS) also shows a significant gap: RWTH-T reaches 0.57, while COTRA-LSCh stands at 0.47 (higher is better). This difference is consistent with the less expressive faces and lower facial stability observed in COTRA-LSCh, and further suggests that the model does not always robustly preserve the signer’s facial features throughout the video. The higher level of variability (standard deviation of 0.098 in COTRA-LSCh) can be interpreted as evidence that identity preservation depends heavily on specific video conditions (e.g., frontal orientation, lighting, face detection), which again points to the less controlled nature of the corpus and the presence of segments with pauses or off-camera glances.

Finally, reference-free sharpness metrics (NR-IQA) provide additional insight: in COTRA-LSCh, some measures suggest that the generated videos may have sharper edges than the originals (e.g., higher Tenengrad values in the generated videos). This finding can be interpreted in two non-mutually exclusive ways: (i) the model is introducing oversharpening artifacts that increase gradients without necessarily improving realism, or (ii) it is compensating for resolution/signal limitations by generating “harder” contours in certain regions. In any case, this result is consistent with a scenario in which structure is preserved, but natural texture is not always faithfully reproduced, and thus a gap is observed between structural and perceptual metrics.

A comparison between RWTH-T and COTRA-LSCh suggests that differences between datasets have a substantial influence on the generator’s error profile. RWTH-T tends to be more homogeneous (with relatively consistent thematic domain, recording context, lighting, and style), which facilitates the learning of fine-grained appearance and stable facial features. COTRA-LSCh, on the other hand, aims to be more diverse and representative: interviewees express themselves more naturally and with individual styles, introducing visual and temporal variability. In this context, it is not surprising that COTRA-LSCh shows competitive results on metrics associated with global coherence (FVD) and structure (SSIM), but faces greater difficulties in fine perceptual quality and identity (FID/LPIPS/IDS). In other words, the model seems to correctly learn “what moves” and “where the body is,” but struggles to stabilize “how critical details look” when the context is more challenging (hands in contact, expressions, and textures).

These results allow us to draw a comprehensive conclusion: the model trained on COTRA-LSCh reliably preserves the spatiotemporal macrostructure of the signal—posture, composition, colors, and body movement—but exhibits systematic weaknesses in perceptual and facial micro-details, especially under occlusion or during hand gestures. In practical terms, this positions COTRA-LSCh as a challenging yet promising case study: the fact that the system achieves an FVD lower than the RWTH-T model and a slightly higher SSIM indicates that the structural foundation is present, and that future improvements should focus on the fidelity of hands and the

face.

As for limitations, previous findings suggest at least three main areas: (i) sensitivity to occlusions and hand movements, which result in distortions and blurring; (ii) low facial expressiveness, which reduces naturalness and may affect linguistic aspects; and (iii) dependence on the distribution of clips used for training (e.g., the presence of pauses or moments of listening), which could bias the model’s prior toward more static faces. Additionally, it is expected that inference hyperparameters such as the guidance scale and the effective time window influence the balance between fidelity to conditions and naturalness of movement; therefore, part of the observed variability could be mitigated through finer domain-specific calibration.

Despite these limitations, the implications of this work are clear. First, it demonstrates the feasibility of adapting a recent conditional video generation methodology (inspired by SignViP) [10] to the context of LSCh, even in a scenario where the corpus exhibits realistic variability and limited capture control. Second, the quantitative and qualitative results suggest that structural conditioning (body and hand poses) effectively anchors the dynamics of signing, providing a solid foundation for advancing toward a full text-to-video translator. In this regard, the central contribution is not only the generator’s performance but also the empirical evidence of which components transfer well and which require reinforcement when operating on a diverse corpus such as COTRA-LSCh.

Looking ahead, the results suggest prioritizing improvements in local quality (hands and face) without compromising temporal consistency. Long-term directions include: expanding the corpus until all recordings are complete; incorporating labels to train primarily on segments where gestures are actually being made (reducing bias toward neutral expressions); exploring resolution adjustments and inference hyperparameters to avoid over-emphasis and artifacts; and, more structurally, building and training the translation module that converts text into discrete states/poses consumable by the diffuser (e.g., a Transformer/GPT-2-type translator to tokens, followed by a decoder that produces embeddings/states), thereby closing the pipeline toward automatic text→video translation in LSCh.

As a final synthesis, the evidence supports a central trade-off in this setting: strong structural conditioning can robustly preserve global spatiotemporal organization even under in-the-wild LSCh variability, but reaching fully natural and linguistically faithful synthesis still depends on improving fine texture and non-manual detail fidelity.

V. CONCLUSION

This work explored the feasibility of adapting a recent conditional video generation methodology inspired by SignViP [10] to the Chilean Sign Language (LSCh) context. Concretely, we implemented and evaluated a signer video generator conditioned on a reference image and structured pose signals, and we assessed its behavior on a newly compiled Chilean corpus (COTRA-LSCh) while also training an analogous model on RWTH-T for comparison. In brief, the

COTRA-LSCh model surpassed its RWTH-T counterpart on FVD and SSIM, while producing visually coherent samples with stable macro-appearance and overall body motion.

Taken together, the contribution of this work is twofold. First, it provides an end-to-end implementation and evaluation of a conditional signer video generator for LSCh, offering initial evidence that the approach transfers beyond highly controlled benchmarks. Second, it delivers an empirical characterization of where the method succeeds (macro-structure, temporal coherence, and appearance anchoring) and where it breaks down (hands under contact/occlusion and facial dynamics), which helps prioritize future model and dataset improvements.

Future work should focus on improving local fidelity without sacrificing temporal consistency. Promising directions include: (i) completing and expanding COTRA-LSCh and leveraging annotations to train primarily on segments where signing is actively performed; (ii) improving hand and face rendering under occlusion via higher-resolution training, stronger inductive biases, and/or dedicated losses for fine details; (iii) refining inference hyperparameters (e.g., guidance and temporal context) to mitigate artifacts such as oversharpening; and (iv) integrating the missing translation component to close the pipeline toward automatic text-to-video generation in LSCh.

We hope these results serve as a solid baseline and a practical stepping stone toward accessible, high-quality sign language video generation for the Chilean deaf community.

ACKNOWLEDGMENT

We thank the LSCh corpus team led by Ana Aguilera, including the annotators and collaborators; in particular, we acknowledge the linguists Fabiola, Paula, Martín, Viviana, and Patricio for their support and contributions to the linguistic annotation process. We also thank the Department of Computer Science at the Faculty of Physical and Mathematical Sciences, University of Chile, for providing the shared computational infrastructure used in this work. Finally, we thank the study participants for their time and contribution to data collection. We also gratefully acknowledge the financial support of ANID through Fondecyt Regular No 1250344.

REFERENCES

- [1] World Health Organization, “World report on hearing,” World Health Organization, Tech. Rep., 2021. [Online]. Available: <https://www.who.int/publications/i/item/world-report-on-hearing>
- [2] Biblioteca del Congreso Nacional de Chile, “Ley 21.303: Modifica la ley n° 20.422 para promover el uso de la lengua de señas,” 2021, publicada el 22 de enero de 2021.
- [3] Ministerio de Desarrollo Social y Familia de Chile, “Presentación de resultados endide 2022: Población adulta,” Observatorio Social, Gobierno de Chile, Tech. Rep., 2023.
- [4] W. C. Stokoe, “Sign language structure: An outline of the visual communication systems of the american deaf,” 1960.
- [5] M. González Moraga and A. Pérez Cuello, “La lengua de señas chilena: un recorrido por su proceso de desarrollo desde una perspectiva multidimensional,” *INES*, no. 47, 2017.
- [6] M. Yáñez and A. Aguilera, “Reconocimiento aislado de la lengua de señas chilena en tiempo real usando deep learning,” *Ingeniare. Revista chilena de ingeniería*, vol. 32, 2024, artículo 29.
- [7] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, “Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus,” in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2012, pp. 3785–3789. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf
- [8] B. Saunders, N. C. Camgoz, and R. Bowden, “Mixed signals: Sign language production via a mixture of motion primitives,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [9] M. Inan, Y. Zhong, S. Hassan, L. Quandt, and M. Alikhani, “Modeling intensification for sign language generation: A computational approach,” *arXiv preprint*, 2022. [Online]. Available: <https://arxiv.org/abs/2203.09679>
- [10] C. Wang, Z. Deng, Z. Jiang, F. Shen, Y. Yin, S. Gan, Z. Cheng, S. Ge, and Q. Gu, “Advanced sign language video generation with compressed and quantized multi-condition tokenization,” *arXiv preprint arXiv:2506.15980*, 2025. [Online]. Available: <https://arxiv.org/abs/2506.15980>
- [11] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. R. Morris, “Sign language recognition, generation, and translation: An interdisciplinary perspective,” in *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 16–31.
- [12] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, and M. Sabokrou, “A survey on recent advances in sign language production,” *Expert Systems with Applications*, vol. 243, p. 122846, 2024. [Online]. Available: <https://doi.org/10.1016/j.eswa.2023.122846>
- [13] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, “Text2sign: Towards sign language production using neural machine translation and generative adversarial networks,” *International Journal of Computer Vision*, vol. 128, pp. 891–908, 2020. [Online]. Available: <https://doi.org/10.1007/s11263-019-01281-2>
- [14] B. Saunders, N. C. Camgoz, and R. Bowden, “Everybody sign now: Translating spoken language to photo realistic sign language video,” *arXiv preprint arXiv:2011.09846*, 2020. [Online]. Available: <https://arxiv.org/abs/2011.09846>
- [15] L. Ventura, A. Duarte, and X. Giró-i Nieto, “Can everybody sign now? exploring sign language video generation from 2d poses,” *arXiv preprint arXiv:2012.10941*, 2020. [Online]. Available: <https://arxiv.org/abs/2012.10941>
- [16] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. 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)*. Nashville, TN, USA: IEEE, 2021, pp. 2735–2744. [Online]. Available: <https://doi.org/10.1109/CVPR46437.2021.00277>
- [17] A. Battisti, “Analysis and assessment of continuous swiss german sign language production of adult learners,” Doctoral dissertation, University of Zurich, Zurich, 2025.
- [18] H. Zhang, R. Shalev-Arkushin, V. Baltatzis, C. Gillis, G. Laput, R. Kushalnagar, L. C. Quandt, L. Findlater, A. Bedri, and C. Lea, “Towards AI-driven sign language generation with non-manual markers,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Yokohama, Japan: ACM, 2025. [Online]. Available: <https://doi.org/10.1145/3706598.3713855>
- [19] R. Pfau and J. Quer, “Nonmanuals: their grammatical and prosodic roles,” in *Sign Languages*, D. Brentari, Ed. Cambridge: Cambridge University Press, 2010, pp. 381–402.
- [20] F. Qi, Y. Duan, H. Zhang, and C. Xu, “Signgen: End-to-end sign language video generation with latent diffusion,” in *Advances in Neural Information Processing*. Springer, 2024, pp. 239–254. [Online]. Available: https://doi.org/10.1007/978-3-031-73668-1_15
- [21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.
- [22] CompVis, “Stable diffusion,” <https://github.com/CompVis/stable-diffusion>, 2022, accessed: 2026-03-22.

- [23] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2022. [Online]. Available: <https://arxiv.org/abs/1312.6114>
- [24] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, “Animatediff: Animate your personalized text-to-image diffusion models without specific tuning,” 2024. [Online]. Available: <https://arxiv.org/abs/2307.04725>
- [25] Z. Yang, A. Zeng, C. Yuan, and Y. Li, “Effective whole-body pose estimation with two-stages distillation,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.15880>
- [26] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.05543>
- [27] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, “Reconstructing hands in 3d with transformers,” 2023. [Online]. Available: <https://arxiv.org/abs/2312.05251>
- [28] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *Image Processing, IEEE Transactions on*, vol. 13, pp. 600 – 612, 05 2004.
- [29] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Towards accurate generative models of video: A new metric & challenges,” 2019. [Online]. Available: <https://arxiv.org/abs/1812.01717>
- [30] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” 2018. [Online]. Available: <https://arxiv.org/abs/1705.07750>
- [31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” 2018. [Online]. Available: <https://arxiv.org/abs/1706.08500>
- [32] G. Parmar, R. Zhang, and J.-Y. Zhu, “On aliased resizing and surprising subtleties in gan evaluation,” in *CVPR*, 2022.
- [33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” 2018. [Online]. Available: <https://arxiv.org/abs/1801.03924>
- [34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, p. 84–90, May 2017. [Online]. Available: <https://doi.org/10.1145/3065386>
- [35] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, p. 5962–5979, Oct. 2022. [Online]. Available: <http://dx.doi.org/10.1109/TPAMI.2021.3087709>
- [36] X. An, J. Deng, J. Guo, Z. Feng, X. Zhu, J. Yang, and T. Liu, “Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc,” in *CVPR*, 2022.
- [37] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou, “Webface260m: A benchmark unveiling the power of million-scale deep face recognition,” in *CVPR*, 2021.