

LSC-CARE: An annotated dataset for medical Catalan Sign Language

Carlos Escolano¹ Joan Gomà Cortés¹ Gemma Barberà²
Gerard Sant³ Maite Melero⁴

¹Universitat Politècnica de Catalunya ²Universitat Pompeu Fabra

³University of Zurich ⁴Barcelona Supercomputing Center

carlos.escolano@upc.edu, joan.goma.cortes@estudiantat.upc.edu,
gemma.barbera@upf.edu, gerard.santmuniesa@uzh.ch, maite.melero@bsc.es

Abstract

Deaf signers face persistent communication barriers in healthcare settings, where reliance on third-party interpreters can compromise privacy precisely when confidentiality matters most. We present LSC-CARE, the first annotated dataset of Catalan Sign Language (LSC) for the medical domain, comprising 2,784 signed segments (approximately 4.5 hours) produced by 11 deaf native signers across ten emergency-care topics. Each segment is paired with time-aligned ID-gloss annotations produced by a deaf native linguist and written Catalan translations produced by an LSC interpreter, following conventions consistent with the LSC SignBank. To assess the dataset's utility for Sign Language Translation, we benchmark three settings: a gloss-free end-to-end video-to-text model, a cascade of sign language recognition and gloss-to-text translation, and an oracle gloss-to-text upper bound. The cascade consistently outperforms the end-to-end approach despite a high gloss error rate, and, unlike it, is statistically invariant to signer identity, while the 21-BLEU gap between cascade and oracle identifies visual sign recognition as the main bottleneck for low-resource medical SLT. We release the dataset to support accessible healthcare technologies for the Catalan Deaf community.

1 Introduction

Catalan Sign Language (*llengua de signes catalana*, LSC) is a low-resource language. It is used in Catalonia, Andorra and the island of Menorca by 25,000 signers in their daily life. Of these, about 12,000 are deaf individuals who use it as their primary language of communication, while the remaining users are hearing individuals, including family members and professionals, such as interpreters, teachers and linguists.¹ In the field of

machine Sign Language Translation (SLT), dominant paradigms based on large volumes of data are not directly applicable to LSC due to a "double minoritization" (Quer, 2012). Despite the existence of LSC linguistic resources (Barberà et al., 2018) and the open-access publication of the first reference corpus of Catalan Sign Language, the Corpus LSC (Institut d'Estudis Catalans, 2025), we still face a shortage of data annotated with the necessary conventions to be interpretable by Artificial Intelligence (AI) systems. Research highlights the complexity of generating photorealistic and linguistically appropriate signed sequences (Bragg et al., 2019); (Rastgoo et al., 2021). To address this gap, approaches are identified that use "core glosses" as an intermediate representation to shift complexity toward the inferential capacity of Large Language Models (LLMs) and preserve translation coherence (Müller et al., 2023).

This paper presents LSC-CARE, the first annotated corpus of Catalan Sign Language for the medical domain, together with the first translation baselines trained on it. Our contributions are threefold. First, we release a corpus of 2,784 signed question-answer segments across ten emergency-care topics, produced by 11 deaf native signers and annotated with time-aligned ID-glosses and written Catalan translations. Second, we benchmark three translation settings — gloss-free end-to-end, gloss-based cascade, and an oracle gloss-to-text upper bound — establishing reference results for medical LSC translation. Third, we provide a statistical analysis of length, topic, and signer effects that characterises where each architecture fails, showing that intermediate gloss supervision yields both higher translation quality and robustness to signer identity under extreme data scarcity. The full dataset can be downloaded from: <https://huggingface.co/>

¹Data sourced from the Catalan Federation for the Deaf (FESOCA), officially cited by the Catalan Secre-

tariat for Language Policy (Generalitat de Catalunya) in <https://tuit.cat/8HN3H>

2 Related Work

Sign language corpora. Progress in SLT has been driven by a small number of annotated benchmarks. RWTH-PHOENIX-Weather 2014T (Camgöz et al., 2018; Forster et al., 2014) pairs roughly 11 hours of German Sign Language weather broadcasts with glosses and German text; CSL-Daily (Huang et al., 2018) provides about 23 hours of Chinese Sign Language across everyday topics; How2Sign (Duarte et al., 2021) offers around 80 hours of instructional American Sign Language. At larger scale, YouTube-SL-25 (Tanzer and Zhang, 2025) mines over 3,000 hours of weakly aligned captioned video across 25+ sign languages, trading annotation quality for volume. A common trait of the annotated corpora is domain restriction: each covers one narrow communicative context, and models trained on them transfer poorly beyond it. For LSC specifically, available resources include the reference Corpus LSC (Barberà et al., 2015; Institut d’Estudis Catalans, 2025), but no resource targets the medical domain, and none provides the gloss-level annotation density required for cascade SLT. LSC-CARE fills this gap at the extreme low-resource end of the landscape: at 4.5 hours it is less than half the size of PHOENIX14T, while covering a domain (emergency healthcare) where translation errors carry direct consequences for patient safety.

Translation architectures. Computational approaches divide into gloss-based and gloss-free paradigms. Gloss-based cascades decompose the task into continuous sign language recognition, typically trained with CTC (Graves et al., 2006), followed by gloss-to-text translation (Camgöz et al., 2018); TwoStream-SLR/SLT (Chen et al., 2022) represents the state of the art, fusing RGB and pose-keypoint streams before CTC decoding. The paradigm benefits from modular supervision and diagnosable errors, but inherits an information ceiling: glosses systematically discard non-manual markers, spatial grammar, and prosody (Müller et al., 2023). Gloss-free end-to-end systems (Zhou et al., 2023) avoid the annotation bottleneck and have become dominant where large data exists (Tanzer and Zhang, 2025; Zhang et al., 2024), but jointly learning visual parsing, alignment, and generation demands training volumes far beyond what low-resource languages can offer. Which paradigm degrades more gracefully under extreme scarcity

remains an open empirical question.

2.1 LSC-CARE: Data Compilation

After doing research into the emergency protocols of the public hospitals in Catalonia, we compiled a set of possible doctor–patient interactions in emergency settings. We recreated possible question-answer pairs between doctor and patient and recorded the LSC version of each sentence. Signer EB is the dedicated question signer, while the remaining 10 signers provide different possible answers to each question. The topics selected were the following: personal data; triage; general; traumatology; gynecology; pediatrics; ophthalmology; otorhinolaryngology; dermatology; pneumology. Table 1 shows the distribution of LSC-CARE by anonymised signer and medical domain.

The recruited deaf signers watched the recorded LSC questions and proposed two or three possible answers to that question. The dataset was therefore augmented in terms of linguistic expressions, lexicon and dialectal variation. This increased the dataset’s linguistic and lexical diversity, as well as its dialectal variation across gender, age, and region. The dataset was professionally recorded in an audiovisual recording studio at the university sign language lab at a resolution of 747x420 or 560x420 pixels and 25 fps (see Figure 1). The studio has a chroma key background to clearly distinguish the signer. Only one signer per shot was recorded.

2.2 Deaf native signers profiles

A total of 11 deaf signers were recorded. They are all native signers, born in deaf families and attended bilingual or specific schools, depending on the age. They are all affiliated to a deaf club and therefore their usual interaction language is LSC. Including native signers allows to describe natural grammatical structures at their optimal stage of neurocognitive development, establishing a reliable baseline against which to compare other developmental variants or instances of second-language acquisition. This confirms that the sample of LSC included in the LSC-CARE contains morphosyntactic structures complex enough to be considered a baseline. Moreover, a balanced amount of gender variation and age groups was achieved, belonging to the three age groups established in the LSC Corpus (Barberà et al., 2015), namely the first group age from 18 to 30 years old, the second age group from 31 to 50 years old and the third age group from 51 to 80 years old. Table 2 summarises the

Signer ID	Dat.	Tri.	Tra.	Gen.	Gyn.	Ped.	Opht.	Oto.	Der.	Pne.	Total
EB(Q)	8	7	37	17	8	12	11	12	21	13	146
BF	-	-	74	-	-	-	-	-	72	-	146
CW	72	-	-	39	-	-	-	41	-	-	152
DC	-	-	-	-	-	-	-	76	71	-	147
DT	19	39	99	46	34	48	54	67	59	30	495
FB	44	22	131	65	23	36	30	37	96	57	541
FW	-	68	-	-	-	70	-	-	-	-	138
GW	-	54	-	-	-	-	-	-	-	75	129
NU	65	45	-	-	51	-	-	-	-	100	261
TT	65	-	244	-	-	-	-	-	109	73	491
YB	43	-	75	-	-	-	-	-	-	20	138
Total	316	235	660	167	116	166	95	233	428	368	2784

Table 1: Number of segments per anonymised signer and medical domain in LSC-CARE. The row *Total* reports the column sum per domain; the column *Total* reports the row sum per signer. [*] The abbreviations for each column are as follows: Dat. = personal data; Tri. = triage; Tra. = traumatology; Gen. = general; Gyn. = gynecology; Ped. = pediatrics; Opht= ophthalmology; Oto. = otorhinolaryngology; Der. = dermatology; Pne. = pneumology.

	Male	Female	Total
18-30 y.o.	BF	CW	2
31-50 y.o.	FB; TT	FW; DC	4
51-80 y.o.	YB; GW	EB; DT; NU	5
Total	5	6	11

Table 2: Signer’s profiles according to age and gender.

different profiles.

2.3 Data characteristics and transcription

The dataset is distributed in Parquet format, with each instance representing a segmented video recording mapped to both its linguistic annotations and target natural language translation. Every sample includes precise video tracking via the signal field—which specifies the local file path to the corresponding .mov or .mp4 video—alongside signal_start and signal_end timestamps in seconds to define the exact temporal boundaries of the segment. For linguistic supervision, the glosses field provides human-annotated sign glosses transcribed in sequential order, while the output field contains the finalized clinical sentence translation in written Catalan.

The glosses used in the annotation process of the LSC-CARE belong to the list of tokens of the LSC database project developed at the Institut d’Estudis Catalans (Quer, 2017) and are part of the project under development LSC SignBank. Using the list of tokens ensures that the glosses used in the annotation remain consistent with the tags. Besides

manually annotating each sign with the corresponding gloss and time slot, the global temporal span of the signed sentence was also defined with the corresponding written Catalan translation. Following usual technical and methodological motivations (Fenlon and Hochgesang, 2022), the ID-glosses annotation of the LSC-CARE was done by a deaf native linguist, while the corresponding Catalan translation was done by an LSC interpreter. This guarantees a correct identification of each sign with an ID-gloss and avoids possible spoken language bias of the transcription of the structure of the signed sentence.

3 Dataset Statistics

The complete dataset contains 2,784 segments totalling approximately 4.5 hours of active signing. Data were partitioned 80/10/10 into training (2,227 segments, 3.5 h), validation (278), and test (279) splits by random shuffling at the segment level with a fixed seed. The splits are therefore signer-dependent: all signers may appear in all splits. This choice is deliberate, given current SLT performance levels on corpora of this size, holding out entire signers would leave too little training signal, but it means reported results reflect seen-signer conditions, a point we return to in the Limitations (Section 9).

The training split contains 26,145 gloss tokens over 3,502 types. The distribution is heavily Zipfian (estimated slope -1.11): the most frequent gloss, the first-person index IX1, alone accounts



Figure 1: Examples of how the dataset was recorded. On the left, the recording setup. On the right, an example from the dataset for the sentence *Yes, I do have diarrhea*.

for 7.0% of all tokens, the 100 most frequent types cover 49.1% of the corpus, while 43.1% of the vocabulary (1,511 types) are singletons and 77.5% of types occur fewer than five times. Mean gloss sequence length is 11.8 glosses per segment (median 10, range 1–66), and mean gloss articulation duration is 0.56 s (10th–90th percentile 0.15–1.07 s).

The annotation convention encodes morphology directly in the label. Plain lexical glosses (e.g. *FEBRE*, *PASTILLA*) comprise 29.7% of types but 54.5% of tokens; hyphenated compound and modified forms — directional indices such as *IX1-BD*, classifier constructions such as *CL-B*, lexicalised compounds such as *TOCAR-COR* — comprise 68.1% of types, of which 83.1% occur fewer than five times, forming most of the long tail. A small set of fingerspelled sequences (75 types, 0.6% of tokens), covering names, medication names, and contact details, is retained as single gloss units following the original annotation. This distribution, a small dense head and a vast sparse tail dominated by productive morphology, mimics an actual scenario where patients may employ infrequent or even out-of-vocabulary expressions.

4 Methodology

We have measured the usefulness of our dataset on the task of Sign Language Translation (SLT) on the following different settings (for detailed hyperparameters, consult section A):

End-to-End: A direct Video-to-Text neural model that translates sign language video directly into target text without an explicit gloss supervision bottleneck.

Cascade: A two-stage pipeline consisting of a Video-to-Gloss recognition model followed by a Gloss-to-Text translation model.

Oracle: A baseline model utilizing ground-truth (reference) glosses as input to the Gloss-to-Text translation model, serving as the empirical upper bound for the cascade architecture.

4.1 End-to-End Architecture

The End-to-End (E2E) pipeline maps an input sequence of raw video frames directly to a target written Catalan sentence without utilizing an intermediate gloss representation (Sant et al., 2025b). Built on top of the MultimodalHugs framework (Sant et al., 2025a) using its video2text module, the model connects a visual feature extractor, a learned multimodal mapper, and a sequence-to-sequence translation backbone.

4.1.1 Visual Feature Extraction

Input sign language videos are temporally subsampled, and each frame is resized to 224×224 pixels and normalized using ImageNet (Deng et al., 2009) statistics. Per-frame spatial representations are extracted using a pretrained CLIP ViT-B/32 visual encoder (Radford et al., 2021). The extracted features are the final Transformer encoder layer’s contrastive projection.

4.1.2 Multimodal Mapping

To bridge the gap between frame-level visual embeddings and the token embedding space expected by the downstream translation decoder, a learned multimodal mapper projects each 512-dimensional visual embedding $x_t \in \mathbb{R}^{d_v}$ at frame t into a mapped feature vector $z_t \in \mathbb{R}^{d_m}$. We employ a linear mapper that projects the visual feature extractor continuous representations to the embedding space of the translation backbone. Different options were explored, but the following simple configuration produced the best results while being the most parameter-wise efficient:

$$z_t = Wx_t + b \quad (1)$$

where $W \in \mathbb{R}^{d_m \times d_v}$ and $b \in \mathbb{R}^{d_m}$.

4.1.3 Translation Backbone

We utilize M2M-100 (418M parameters) (Fan et al., 2021)² as the translation backbone. The encoder input embeddings are replaced directly by the mapper’s projected visual frame embeddings, while the decoder retains its pretrained multilingual sequence-to-sequence weights. Translation direction is explicitly prompted via special control tokens: an `__lsc__` token is prepended to the input sequence to indicate a sign language source, and the decoder is initialized with `__ca__` to force generation in written Catalan.

4.2 Cascade Architecture

The Cascade pipeline decomposes sign language translation into two modular stages: Sign Language Recognition (SLR / Video-to-Gloss) followed by Gloss-to-Text (G2T) translation (Camgöz et al., 2018).

4.2.1 Stage 1: Sign Language Recognition (Video-to-Gloss)

The SLR model extracts per-frame visual features using CLIP ViT-B/32. Because per-frame CLIP embeddings are temporally independent, temporal context is introduced through two stacked sequence-modeling components:

Temporal Convolutional Network (TCN): A two-layer 1D convolutional network processes the feature sequence along the time axis. The first layer subsamples the temporal dimension while projecting features into a higher dimension; the second layer applies a dilated convolution to extend the effective receptive field without further downsampling.

Temporal Transformer Encoder: The downsampled sequence is processed by a stack of Transformer encoder layers (Vaswani et al., 2017). Standard sinusoidal positional encodings are added before the first layer to inject temporal order:

$$\begin{aligned} \text{PE}_{(t,2i)} &= \sin\left(\frac{t}{10000^{2i/d}}\right), \\ \text{PE}_{(t,2i+1)} &= \cos\left(\frac{t}{10000^{2i/d}}\right) \end{aligned} \quad (2)$$

where d is the model dimension and t is the frame position. Pre-layer normalization is applied before each sub-layer for training stability.

²https://huggingface.co/facebook/m2m100_418M

A linear projection layer maps the output of the Temporal Transformer encoder to a frame-level distribution over the target gloss vocabulary (plus the CTC blank token ϵ). The recognition stage is trained alignment-free using Connectionist Temporal Classification (CTC) loss. At inference time, predicted frame alignments are collapsed into a discrete gloss sequence g via greedy or beam-search decoding.

4.2.2 Stage 2: Gloss-to-Text Translation

The predicted gloss sequence g is passed to the downstream G2T model, which consists of a fine-tuned M2M-100 (418M) backbone trained on (gloss sequence, Catalan sentence) pairs. The same control token protocol is maintained: `__lsc__` prompts the input gloss sequence and `__ca__` triggers target text generation.

4.3 Oracle Baseline

Because the Cascade pipeline decouples recognition from translation, we also evaluate the G2T model in an **Oracle** setting. In this configuration, the visual SLR model is bypassed entirely, and ground-truth (reference) glosses from the dataset are fed directly into the fine-tuned M2M-100 translation backbone. This setup isolates the Gloss-to-Text generation capacity from visual perception noise, establishing an empirical performance upper bound for the cascade paradigm.

4.4 Evaluation

Evaluation is conducted using standard lexical precision metrics (BLEU (Papineni et al., 2002), ChrF (Popović, 2015)), TER (Snover et al., 2006) and semantic/quality estimation metrics (BERTScore F1 (Zhang et al., 2020), CometKiwi (Rei et al., 2022)). Statistical significance is assessed via Pearson correlation (r), Spearman rank correlation (ρ), one-way ANOVA (Stuhle and Wold, 1989), and Kruskal-Wallis non-parametric tests (Ostertagova et al., 2014).

5 Experimental Setup

All models and experimental pipelines were implemented in Python using PyTorch (Paszke et al., 2019) as the primary deep learning framework. Evaluation was performed using sacrebleu (Post, 2018)³. All experiments were conducted on a single NVIDIA RTX 3090 with 24GB of VRAM

³"nrefs:1|case:mixed|eff:nltk:13|smooth:expl|version:2.5.1"

and using 50GB of system RAM. Each training took approximately 2 days under these settings.

Due to structural differences in model architecture and infrastructure requirements, the two core translation paradigms were implemented using distinct framework ecosystems.

5.1 End-to-End Pipeline Implementation

The End-to-End (e2e) pipeline was built using `multimodalhugs`, an open-source extension of the Hugging Face ecosystem designed for training, evaluating, and deploying multimodal architectures within standard Hugging Face workflows. The framework provides a generalized data pipeline capable of processing heterogeneous multimodal inputs. Experiment execution is managed via YAML-based configuration files, with full integration with Weights & Biases (W&B) for real-time tracking, metric logging, checkpoint saving, and experiment resumption.

5.2 Gloss-Based Cascade Pipeline Implementation

Because existing open-source libraries provided insufficient support for the two-stage visual-to-gloss-to-text pipeline, the **Cascade** architecture was implemented from scratch using PyTorch:

Video-to-Gloss (V2G): The first-stage visual recognition network uses a custom PyTorch implementation combining the feature extraction and projection modules described in Section 4.2.

Gloss-to-Text (G2T): The second-stage text generation module directly fine-tunes a pre-trained M2M-100 (418M) model checkpoint using the native Hugging Face Trainer API.

Both stages of the gloss-based pipeline support metric-driven early stopping, automatic checkpointing, and W&B experiment tracking.

6 Results

In this section, we present an empirical evaluation of three translation setups: the direct End-to-End (**E2E**) model, the two-stage **Cascade** model, and the **Oracle** model (which consumes ground-truth target glosses to establish an empirical upper bound). We analyze corpus-level performance, sequence length behaviors, domain topic sensitivity, and demographic fairness across signers, age groups, and genders.

6.1 Overall System Performance and Output Lengths

Table 3 summarizes overall corpus-level performance alongside generated sentence length statistics across all three architectures. Ground-truth gloss translation (**Oracle**) outperforms all models across every metric, reaching a BLEU score of 25.97, ChrF of 48.14, BERTScore F1 of 84.28, and CometKiwi of 55.86.

The two-stage **Cascade** model consistently outperforms direct **E2E** translation on surface and lexical precision, gaining +1.90 BLEU points, +4.19 ChrF points, and +3.68 BERTScore F1 points. However, both architectures yield comparable overall semantic quality according to CometKiwi (42.40 for E2E vs. 41.58 for Cascade).

A notable difference emerges in generated text volume:

E2E model: generates an average prediction length of 13.22 ± 9.43 words, closely mirroring the target reference length distribution (15.09 ± 9.73 words).

Cascade model: exhibits severe sentence truncation, averaging only 9.67 ± 6.80 predicted words per sample.

Comparing **Cascade** against **Oracle** isolates the performance bottleneck introduced during visual sign recognition (Video-to-Gloss). Visual recognition errors account for large drops across all metrics: -21.17 BLEU ($25.97 \rightarrow 4.80$), -28.48 ChrF ($48.14 \rightarrow 19.66$), -11.15 BERTScore F1 ($84.28 \rightarrow 73.13$), and -14.28 CometKiwi ($55.86 \rightarrow 41.58$). Furthermore, removing video recognition errors restores prediction sequence length from 9.67 words to 14.56 ± 10.24 words in the Oracle setup.

6.2 Length Effects and Correlation Analysis

We evaluated the relationship between sequence length variables and performance metrics using sample-level Pearson (r) and Spearman (ρ) correlation tests (Table 8).

All models display statistically significant negative correlations between target lengths (Gloss Length and Reference Word Length) and n -gram precision (BLEU, $p < 0.001$). Longer input signing sequences compound translation degradation.

However, output length dynamics differ fundamentally:

Model	Avg Words	WER ↓	BLEU ↑	ChrF ↑	BS F1 ↑	CKiwi ↑
E2E	13.22 ± 9.43	–	2.90	15.47	69.45	42.40
Cascade	9.67 ± 6.80	83.24	4.80	19.66	73.13	41.58
Oracle	14.56 ± 10.24	–	25.97	48.14	84.28	55.86

Table 3: Overall Corpus Results and Prediction Length Statistics across Evaluation Pipelines (Reference Target: 15.09 ± 9.73 words, 11.04 ± 6.77 glosses).

E2E Character Overlap: In E2E, generating longer sequences strongly correlates with higher character-level precision (ChrF Pearson $r = 0.320, p < 0.001$; Spearman $\rho = 0.631, p < 0.001$). While producing longer hypotheses increases character match rates, it does not translate into higher sequence-level BLEU or BERTScore F1.

Oracle Uniform Degradation: In the Oracle system, longer generated sequences yield uniformly lower scores across all metrics ($p < 0.05$), suggesting that longer sequences remain harder to translate even when visual recognition noise is removed.

6.3 Categorical, Demographic, and Domain Analysis

To analyze system stability across topics and demographics (Signer ID, Age Group, Gender), we conducted parametric one-way ANOVA (F) and non-parametric Kruskal-Wallis (H) hypothesis tests (Table 4).

The statistical test comparisons yield three key structural insights:

Signer and Demographic Invariance in Cascade: In the **Cascade** model, performance variations across signers, age groups, and genders are statistically non-significant ($p > 0.05$). This is consistent with the intermediate gloss sequence abstracting away signer-specific visual variation, although, given the low absolute scores, part of this invariance may reflect a floor effect (see Section 9).

Demographic Sensitivity in E2E: The **E2E** architecture exhibits significant sensitivity across signers ($p < 0.001$), age groups ($p < 0.01$), and genders ($p < 0.05$). For example, the 18–30 age bracket achieves higher BLEU (2.72) and BERTScore F1 (69.24) than older signers, while female signers score higher on CometKiwi (28.07 vs. 24.41). E2E models appear to directly inherit visual domain bias without intermediate abstraction.

Origin of Demographic Disparities: In the **Oracle** system, demographic variance across Age Group and Gender drops to complete statistical non-significance ($p > 0.05$). This suggests that observed demographic disparities in sign language translation originate entirely within the visual feature extraction phase rather than downstream text translation.

7 Discussion

The comparative evaluation yields a central empirical finding: an explicit gloss bottleneck appears to be a useful inductive bias in an extreme low-resource Sign Language Translation (SLT) setting. The cascade pipeline outperforms the end-to-end model on overlap-based metrics, with comparable CometKiwi scores, despite high error rates in the initial recognition stage. This performance differential indicates that even noisy symbolic intermediates retain sufficient structural information to constrain sequence generation better than direct visual conditioning, which—when trained on approximately 3.5 hours of video—exhibits severe target-language degeneration toward fluent yet unfaithful output. Qualitatively, the primary failure mode of the end-to-end architecture is topic hallucination, characterized by syntactically fluent Catalan output detached from source content, occasionally degenerating into long templated narratives. In contrast, cascade errors are traceable directly to transcription noise within the intermediate recognition stage. Section C shows a list of examples of the most common translations errors produced by our systems.

Detailed analysis of generation errors via Translation Edit Rate (TER) alignment—which accounts for phrasal block shifts alongside standard edit distance—identifies 3,872 edit operations across 4,210 reference tokens, yielding an overall cascade TER of 91.97%. The error distribution is dominated by substitutions (43.87%, 1,847 instances) and deletions (40.02%, 1,685 instances), whereas insertions

Category	Metric	Cascade		End-to-End		Oracle	
		ANOVA F	Kruskal H	ANOVA F	Kruskal H	ANOVA F	Kruskal H
Topic	BLEU	1.420 (n.s.)	14.385 (n.s.)	0.421 (n.s.)	14.544 (n.s.)	1.597 (n.s.)	18.404*
	ChrF	1.813 (n.s.)	15.651 (n.s.)	0.434 (n.s.)	11.503 (n.s.)	3.001**	22.577**
	CometKiwi	1.981*	15.892*	5.444***	42.852***	3.053**	21.716**
Signer	BLEU	0.891 (n.s.)	15.283 (n.s.)	2.979**	31.874***	1.238 (n.s.)	15.154 (n.s.)
	ChrF	0.823 (n.s.)	2.907 (n.s.)	3.724***	60.757***	1.404 (n.s.)	14.714 (n.s.)
	CometKiwi	1.655 (n.s.)	13.806 (n.s.)	5.079***	47.073***	2.520**	22.073*
Age Group	BLEU	0.607 (n.s.)	2.945 (n.s.)	9.974***	11.930**	0.052 (n.s.)	0.217 (n.s.)
	CometKiwi	1.465 (n.s.)	2.303 (n.s.)	5.819**	11.868**	1.549 (n.s.)	3.361 (n.s.)
Gender	CometKiwi	3.292 (n.s.)	0.996 (n.s.)	5.887*	3.969*	0.141 (n.s.)	0.181 (n.s.)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. = not significant.

Table 4: ANOVA (F) and Kruskal-Wallis (H) Tests Across Domain Topics and Demographic Factors.

(4.13%, 174 instances) and block shifts (3.94%, 166 instances) occur at markedly lower frequencies. The low shift frequency indicates that word-order misalignment is minimal, indicating that phrase structure is largely preserved when target vocabulary is correctly generated. Conversely, high deletion rates reflect a sequence collapse associated with the underlying Connectionist Temporal Classification (CTC) recognition bottleneck in visually ambiguous regions. High substitution rates point to downstream language model hallucination when generating text from truncated gloss input. Performance varies across thematic domains, with higher error rates recorded in specialized categories such as Dermatology (96.67% TER) and Pneumology (94.24% TER) compared to broader categories like Triage (87.99% TER) and Personal Data (89.29% TER). Errors on fingerspelled and numeric content (e.g., dosages, contact details) remain clinically critical while contributing proportionally little to aggregate overlap metrics.

Categorical decomposition across signers reveals bounded variance (ranging from 82.67% TER for signer GW to 101.11% for CW), consistent with the hypothesis that symbolic gloss representations mitigate signer-specific visual variance. However, given the high baseline error rate, this relative invariance may partially reflect a floor effect. Furthermore, as data splits are signer-dependent, generalization to unseen signers remains unverified. Finally, oracle experiments delineate the upper bound of gloss-mediated approaches: even with error-free recognition, a subset of target sentences receives low scores, highlighting information unencoded by standard gloss sequences, such as spatial grammar, non-manual markers, and deictic references.

These findings point toward two clear research directions: improving visual recognition through motion-aware encoders and pose dynamics, and enriching the intermediate representation beyond unigram glosses without removing the structured intermediate supervision required in data-constrained settings.

8 Conclusions

We presented LSC-CARE, the first annotated corpus of Catalan Sign Language for medical communication: 2,784 segments across ten emergency-care topics, signed by 11 deaf native signers and annotated with ID-glosses and Catalan translations by deaf and hearing experts. Benchmarking three translation settings, we found that a gloss-based cascade clearly outperforms a gloss-free end-to-end model under extreme data scarcity, that the cascade is robust to signer identity where the end-to-end model is not, and that the dominant bottleneck is recall in the visual recognition stage rather than gloss-to-text generation. These results establish reproducible baselines for medical LSC translation and suggest that for severely under-resourced sign languages, structured intermediate supervision buys more than its annotation cost. Future work includes motion- and pose-aware visual encoders targeting classifier constructions and dedicated letter-level fingerspelling recognition. LSC-CARE and evaluation code will be publicly released.

9 Limitations

LSC-CARE is a domain-specific resource, and its limitations follow from its design. First, the data consists of scripted question-answer interac-

tions recreated from hospital emergency protocols rather than spontaneous doctor–patient conversation; elicited signing may differ from naturalistic clinical discourse in register and turn-taking, and models trained on it may not transfer to conversational or non-medical settings.

Second, our data splits are random at the segment level, so all signers and topics are seen during training. We consider signer-independent evaluation premature at current performance levels and corpus size, but this means our results characterise seen-signer conditions only, and the cascade’s signer invariance should not be read as generalisation to unseen signers.

Third, the corpus is small (4.5 hours), constraining architectural conclusions: the low absolute scores mean some null results, including aspects of the signer-invariance finding, may partly reflect floor effects. Our correlation and categorical analyses also involve many hypothesis tests without multiple-comparison correction, so individual significance values should be interpreted cautiously.

Fourth, sequences making heavy use of classifier predicates were excluded from the current annotated release because lexicon-based conventions cannot yet transcribe them consistently. Since classifier predicates are central to symptom description, this removes some of the most clinically relevant material; annotating it is a priority for future releases. Relatedly, fingerspelled items are retained as single opaque gloss units, which our error analysis identifies as a weakness for exactly the content where errors matter most, such as medication names and identifiers.

Finally, evaluation is automatic only. No human evaluation with Deaf users, interpreters, or clinicians was conducted, and automatic metrics under-penalise clinically critical errors such as laterality mistakes or corrupted contact information; the systems reported here are research baselines, not deployable clinical tools.

10 Ethical Statement

The procedure performed in this study was conducted in accordance with the ethical standards of the institutional research committee of our institution. Written informed consent was obtained from all 11 participating signers prior to recording, covering both their participation in the study and the public release of their identifiable video recordings as part of the dataset; participants were informed

that, unlike text or audio, signed video cannot be anonymised without destroying its linguistic content, and retained the right to withdraw at any point before release.

Regarding representativeness, the signer pool was balanced by age and gender (Table 2) but is homogeneous in ethnicity, with all 11 signers being white/Caucasian. This reflects the demographic composition of the Catalan Deaf signing community from which participants were recruited, rather than a sampling choice, but it means the dataset does not capture the visual or linguistic variation associated with more ethnically diverse Deaf communities, and systems trained on it should not be assumed to generalise to signers outside this demographic profile without further validation.

Acknowledgments

This work/research has been promoted and financed by the Government of Catalonia through the Aina Project. This work is also supported by MLLM4TRA (PID2024-158157OB-C32) funded by MCIN/AEI/10.13039/501100011033/FEDER, UE. Carlos Escolano is funded by Grant LLM4SL (PID2024-157855OA-C33) MICI-U/AEI/10.13039/501100011033/FEDER, UE. Gerard Sant is funded by the SIGMA project (grant no. G-95017-01-07), supported by the Digital Society Initiative (DSI) at the University of Zurich. We warmly thank Delfina Aliaga and Mireia Isal for their work on the annotation and translation processes, as well as the deaf LSC participants who participated in the video recordings, without whom this research would not have been possible.

References

- Gemma Barberà, Eulàlia Ribera, and Cristina Gelpí. 2018. *Recursos lingüístics i educatius de la llengua de signes catalana (LSC)*. *Llengua, Societat i Comunicació*, 16:16–25.
- Gemma Barberà, Josep Quer, and Santiago Frigola. 2015. *Primers passos cap a la documentació de discurs signat. El projecte pilot de constitució del corpus de la llengua de signes catalana* [First steps towards the documentation of signed discourse. The pilot project for the creation of the Catalan Sign Language corpus]. *Treballs de Sociolingüística Catalana*, 25:287–302.
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreaux, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa

- Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. [Sign language recognition, generation, and translation: An interdisciplinary perspective](#). In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '19*, page 16–31, New York, NY, USA. Association for Computing Machinery.
- Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. [Neural sign language translation](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society.
- Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. [Two-stream network for sign language recognition and translation](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [Imagenet: A large-scale hierarchical image database](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i-Nieto. 2021. [How2sign: A large-scale multimodal dataset for continuous american sign language](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021*, pages 2735–2744. Computer Vision Foundation / IEEE.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *J. Mach. Learn. Res.*, 22:107:1–107:48.
- Jordan Fenlon and Julie A. Hochgesang, editors. 2022. *Signed Language Corpora*. Gallaudet University Press.
- Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. [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 2014, Reykjavik, Iceland, May 26–31, 2014*, pages 1911–1916. European Language Resources Association (ELRA).
- Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. [Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks](#). In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25–29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.
- Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. [Video-based sign language recognition without temporal segmentation](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018*, pages 2257–2264. AAAI Press.
- Institut d’Estudis Catalans. 2025. [Corpus de referència de la llengua de signes catalana \(LSC\) \(CORPUS LSC\)](#).
- Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. [Considerations for meaningful sign language machine translation based on glosses](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.
- Eva Ostertagova, Oskar Ostertag, and Jozef Kováč. 2014. [Methodology and application of the kruskal-wallis test](#). *Applied Mechanics and Materials*, 611:115–120.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [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. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Josep Quer. 2012. Legal pathways to the recognition of sign languages: A comparison of the catalan and

- spanish sign language acts. *Sign Language Studies*, 12(4):565–582.
- Josep Quer. 2017. [Els projectes de codificació de la llengua de signes catalana \(LSC\)](#). *Estudis Romànics*, 39:445–452.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. [Sign language recognition: A deep survey](#). *Expert Syst. Appl.*, 164:113794.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Gerard Sant, Zifan Jiang, Carlos Escolano, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2025a. [Multimodalhugs: Enabling sign language processing in hugging face](#). *CoRR*, abs/2509.09729.
- Gerard Sant, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2025b. [Modality matters: Training and tokenization effects in sign-to-text translation](#). In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents, IVA Adjunct 2025, Berlin, Germany, September 16-19, 2025*, pages 6:1–6:8. ACM.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Lars Ståhle and Svante Wold. 1989. [Analysis of variance \(anova\)](#). *Chemometrics and Intelligent Laboratory Systems*, 6(4):259–272.
- Garrett Tanzer and Biao Zhang. 2025. [Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. [Scaling sign language translation](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. [Gloss-free sign language translation: Improving from visual-language pretraining](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20814–20824. IEEE.

A Training Hyperparameters

Hyperparameter	Value
Trainable components	All model components
Projected feature dimension	512
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.998$)
LR schedule	Inverse square root
Effective batch size	64 (1×64 accumulation steps)
Gradient clipping	Disabled (max grad norm = 0)
Weight decay	0
Label smoothing	0.1
Frame stride	2
Precision	FP16
Early stopping	Patience = 15 (metric: SacreBLEU)
Decoding metric	SacreBLEU + ChrF

Table 5: Hyperparameter configurations for the End-to-End (E2E) translation model.

B Length and Performance Comparison

Component	Hyperparameter	Value
<i>Visual Backbone (shared)</i>		
CLIP encoder	Checkpoint	openai/clip-vit-base-patch32
CLIP encoder	Output representation	512-dim contrastive projection
<i>Temporal CNN (Video-to-Gloss)</i>		
Conv layer 1	Filters	2,048
Conv layer 1	Kernel size	5
Conv layer 1	Stride	2
Conv layer 1	Padding	Same-style
Conv layer 2	Filters	512
Conv layer 2	Kernel size	5
Conv layer 2	Dilation	2 (effective receptive field: 9 frames)
Conv layer 2	Stride	1
Conv layer 2	Output dimension	512
Both layers	Normalisation	Batch normalisation
Both layers	Activation	ReLU
Both layers	Dropout	0.2
<i>Temporal Transformer Encoder (Video-to-Gloss)</i>		
Positional encoding	Type	Sinusoidal
Positional encoding	Dropout	0.1
Encoder	Layers	6
Encoder	Model dimension	512
Encoder	Attention heads	8
Encoder	FFN dimension	2,048
Encoder	Attention dropout	0.1
Encoder	FFN dropout	0.1
<i>Output Head (Video-to-Gloss)</i>		
Linear head	Input dimension	512
Linear head	Gloss vocabulary size	3,951 classes
Linear head	Training/inference loss	CTC
<i>Gloss-to-Text (G2T)</i>		
Backbone	Architecture	M2M-100
Backbone	Parameters	418M
Backbone	Shared vocabulary	128,113 tokens
Backbone	Encoder layers	12
Backbone	Decoder layers	12
Backbone	Model dimension	1,024
Backbone	FFN dimension	4,096
Backbone	Special token __lsc__	Marks input as LSC gloss sequence

Table 6: Hyperparameter configuration for the Cascade architecture using the 512-dimensional CLIP contrastive projection variant, split into the Video-to-Gloss (V2G) recognition stage and the Gloss-to-Text (G2T) translation stage.

Hyperparameter	Video-to-Gloss (V2G)	Gloss-to-Text (G2T)
Base model / backbone	CLIP ViT-B/32 (unfrozen)	M2M-100 418M (fine-tuned)
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.998$)	AdamW ($\beta_1 = 0.9, \beta_2 = 0.998$)
Learning rate (backbone)	5×10^{-6}	—
Learning rate (other / full)	5×10^{-4}	3×10^{-5}
LR schedule	Linear warmup + cosine decay	Warmup ratio 0.1
Warmup steps / ratio	200 steps	Ratio 0.1
Max epochs	80	15
Batch size (per device)	2	8
Gradient accumulation	32 (effective 64)	4
Weight decay	0.01 (backbone) / 0.05 (full)	0.05
Gradient clipping	5.0	—
Frame stride	2	—
Max sequence length	—	128 tokens
Decoding	—	Beam search ($k = 5$)
Precision	FP16	FP16
Early stopping	Patience 10 (WER)	Patience 5 (SacreBLEU)

Table 7: Hyperparameter configurations for the Cascade model, split in Video-to-Gloss (V2G) and Gloss-to-Text (G2T) model training.

Variable	Metric	Cascade		End-to-End		Oracle	
		Pearson r	Spearman ρ	Pearson r	Spearman ρ	Pearson r	Spearman ρ
Gloss Length	BLEU	−0.266***	−0.422***	−0.187**	−0.398***	−0.277***	−0.233***
	ChrF	−0.095 (n.s.)	0.036 (n.s.)	−0.044 (n.s.)	−0.115 (n.s.)	−0.238**	−0.184**
	BERTScore F1	−0.169**	−0.131*	−0.030 (n.s.)	0.006 (n.s.)	−0.320***	−0.273***
	CometKiwi	−0.113 (n.s.)	−0.017 (n.s.)	−0.003 (n.s.)	0.051 (n.s.)	−0.190**	−0.133*
Ref Word Length	BLEU	−0.309***	−0.512***	−0.242***	−0.473***	−0.303***	−0.251***
	ChrF	−0.143*	0.007 (n.s.)	−0.098 (n.s.)	−0.156*	−0.321***	−0.303***
	BERTScore F1	−0.168**	−0.119 (n.s.)	0.023 (n.s.)	0.091 (n.s.)	−0.335***	−0.295***
	CometKiwi	−0.108 (n.s.)	−0.049 (n.s.)	0.005 (n.s.)	0.075 (n.s.)	−0.198**	−0.149*
Pred Word Length	BLEU	−0.161**	−0.166**	0.006 (n.s.)	0.063 (n.s.)	−0.266**	−0.183**
	ChrF	0.066 (n.s.)	0.280***	0.320***	0.631***	−0.228*	−0.159*
	BERTScore F1	−0.069 (n.s.)	−0.008 (n.s.)	−0.052 (n.s.)	0.040 (n.s.)	−0.345***	−0.253***
	CometKiwi	−0.102 (n.s.)	−0.032 (n.s.)	−0.059 (n.s.)	−0.013 (n.s.)	−0.224**	−0.155*

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. = not significant.

Table 8: Sample-Level Correlation Analysis between Length Variables and Evaluation Metrics.

C Translation Examples

The examples below focus on the most frequent and clinically important categories: catastrophic hallucination, V2G recognition collapse, partial gloss confusion, structured contact-information errors, and templated over-generation. **Green highlighting** marks content that correctly matches or preserves the reference, while **red highlighting** marks the specific error being illustrated.

Topic hallucination. This is the most common error of the gloss-free end-to-end system. Rather than producing a disfluent or partially correct translation, the model frequently generates a fluent, grammatical Catalan sentence about an entirely different clinical topic. [Example C.1](#) illustrates this behaviour.

Example C.1: Topic substitution in the gloss-free system (Sample 1)

Reference gloss: SI IX1 PARE-2 ULLS-1 UN IX-ULLS-DRETA-1 CATARACTES-4 UN

Reference: Sí, el meu pare té cataractes a l’ull dret.

(“Yes, my father has cataracts in his right eye.”)

E2E prediction: Sí, quan vaig amb bicicleta i em ve l’aire de cara em ploren els ulls, sí.

(“Yes, when I ride my bike and the wind blows in my face my eyes water, yes.”)

Cascade: Sí, el meu pare nom és el meu pare. (BLEU-1 55.6, chrF 0.712)

Oracle: Sí, el meu pare té un ull esquerre amb cataractes, un altre amb el dret. (BLEU-1 46.7, chrF 0.730)

The end-to-end system retains only the discourse marker “Sí” and replaces the clinical statement with an unrelated but fluent sentence about watery eyes while cycling. The output remains loosely eye-related, but it does not contain any of the source information about the father, cataracts, or the affected right eye. The cascade hallucinates less because the recognised glosses still preserve the subject, but the missing cataract and eye-location signs lead the G2T model to produce a semantically incorrect sentence. The oracle translation shows a different problem: even with accurate glosses, the text generator can distort laterality, which is clinically significant in medical communication.

Recognition-stage collapse masking a correct G2T stage. A recurring cascade pattern is that the G2T stage performs well when given an accurate gloss sequence, but the full pipeline fails because the upstream V2G model produces an empty or severely corrupted recognition. [Example C.2](#) is the clearest case: the oracle exactly reproduces the reference, while the cascade collapses to a generic filler utterance.

Example C.2: V2G recognition collapse (Sample 35)

Reference gloss: DORMIR PROBLEMA HAVER-NO HAVER-NO RES-5 RES-5

Predicted gloss: (empty)

Reference: No tinc cap problema per dormir. (“I have no problem sleeping.”)

E2E prediction: Em fa mal el cul, a les vertèbres de baix.

(“My bottom hurts, in the lower vertebrae.”)

Cascade: Sí, sí. (BLEU-1 0.0, chrF 0.103)

Oracle: No tinc cap problema per dormir. (BLEU-1 100.0, chrF 1.000, exact match)

Here the V2G model recognises no glosses at all for a six-token reference sequence. The G2T model therefore receives no meaningful source content and falls back to “Sí, sí.”, producing simultaneous hallucination, omission, and under-generation. The end-to-end system also fails, but in a different way: it hallucinates an unrelated complaint about lower-back pain.

[Example C.3](#) shows a less extreme version of the same phenomenon. The V2G output preserves the falling/stairs predicate, but substitutes the shopping-centre context with unrelated signs, causing the final translation to combine a correct event frame with invented details.

Example C.3: Partial gloss confusion (Sample 97)

Reference gloss: IX1-5 CENTRE COMERCIAL CL-2D BAIXANT-ESCALA CL-2 PERSONA-CAIENT-A-LA-VOLTA

Predicted gloss: PATINET DUTXAR-SE IX-LLOC ESCALA CL-B CL-2 PERSONA-CAIENT

Reference: Vaig caure per les escales del centre comercial.

(“I fell down the stairs of the shopping centre.”)

E2E prediction: Em fa mal el dit gros del peu esquerre, perquè em vaig donar un cop al peu amb la vorera.

(“My big toe hurts, because I hit my foot on the curb.”)

Cascade: M’he dutxat amb patinet i he caigut per les escales. (BLEU-1 20.0, chrF 0.720)

Oracle: Vaig caure per les escales del centre comercial. (BLEU-1 100.0, chrF 1.000, exact match)

The recognised gloss sequence keeps PERSONA-CAIENT and ESCALA, so the cascade correctly produces the notion of falling on stairs. However, CENTRE COMERCIAL is replaced by PATINET DUTXAR-SE, leading to a fluent but impossible clause about showering with a scooter. This type of output is particularly risky because it can remain plausible as a medical complaint while being factually wrong.

Numeric and contact-information errors. Personal contact information is encoded through fingerspelling and digit signs, placing strong demands on temporal resolution and fine-grained handshake recognition. [Example C.4](#) shows how a small recognition error can have a large semantic cost.

Example C.4: Fingerspelled email address (Sample 2)

Reference gloss: IX1 MAIL VINT I VUIT DC A-N-N-A CINC ARROVA HOTMAIL DC C-O-M

Predicted gloss: IX1 MAIL DC A-N-N-A CINC DC PUNT-ØE DC C-O-M

Reference: El meu correu és vint-i-vuit anna cinc arrova hotmail com.

(“My email is twenty-eight anna five at hotmail com.”)

E2E prediction: El meu correu és EAP, ara no ho sé, EAP asterisc al mig, ni a dalt ni a baix, al mig, vint-i-quatre arrova gmail com.

Cascade: El meu correu és anna cinc punt com. (BLEU-1 70.0, chrF 0.745)

Oracle: El meu correu és vint-i-vuit, anna, cinc, arrova hotmail com. (BLEU-1 70.0, chrF 0.975)

The V2G model drops the leading numeral sequence VINT I VUIT and misrepresents the domain marker, so the cascade preserves the broad email template while losing crucial identifier con-

tent. The BLEU-1 score remains relatively high because many surrounding words overlap, but a missing two-digit element in contact information is much more serious than the metric suggests.

Templated hallucination and over-generation.

The most severe over-generation cases reveal a distinct failure mode in which the end-to-end model appears to fall back on a long memorised response largely independent of the input. [Example C.5](#) shows the most extreme instance.

Example C.5: Templated over-generation (Sample 40)

Reference gloss: IX-CARA-5 SI

Reference: Sí, uf.

(“Yes, ugh.”)

E2E prediction: Sí, em van aplicar calor per poder fer els estiraments, em van aplicar aquesta calor als braços i cames. Però quan vaig començar a jugar a futbol, estava corrent i vaig notar que se m’enrampava la cama, em feia mal i vaig agafar gel, quan vaig aplicar el fred ja em va marxar la rampa, però com que em costava caminar al final vaig seure per descansar.

(A 56-word fabricated narrative about sports injury, heat therapy, and cramping.)

Cascade: Sí, sí. (BLEU-1 50.0, chrF 0.714)

Oracle: Sí, a la cara. (BLEU-1 25.0, chrF 0.476)