

APPENDIX

Lost in Translation, Found in Context: Sign Language Translation with Contextual Cues

Youngjoon Jang^{*1,2} Haran Raajesh^{*3,4} Liliane Momeni¹ Gül Varol^{1,3} Andrew Zisserman¹

¹Visual Geometry Group, University of Oxford, UK

²KAIST, Daejeon, Republic of Korea

³LIGM, École des Ponts, IP Paris, Univ Gustave Eiffel, CNRS, France

⁴CVIT, IIT Hyderabad, India

<https://www.robots.ox.ac.uk/~vgg/research/litfic/>

* denotes equal contribution

This appendix supplements the main paper by providing additional implementation details (Appendix A), experiments (Appendix B), and qualitative results (Appendix C).

A Implementation Details	1
A.1. LLM Evaluation metric	1
A.2. Prompt details	1
A.3. Background description collation	2
A.4. BOBSL mapping network	3
A.5. How2Sign training details	3
B Additional Experiments	4
B.1. Llama decoder variants	4
B.2. Combining different cues	4
B.3. Missing cue scenario	5
B.4. Number of previous sentences	5
B.5. Background frame sampling rate	5
B.6. ISLR performance on How2Sign	6
B.7. Reproducing GFSLT and Sign2GPT on PHOENIX14T	6
B.8. Evaluation of our model on PHOENIX14T . .	6
C Additional Qualitative Results	7

A. Implementation Details

We provide details on the LLM evaluation metric (Appendix A.1), the list of prompts for the translation model (Appendix A.2), the background description collation (Appendix A.3), the architectural design for the mapping network on BOBSL (Appendix A.4), and the training procedure on How2Sign (Appendix A.5).

A.1. LLM Evaluation metric

As mentioned in Sec. 4.1, our LLM-based evaluation metric is adapted from the CLAIR framework [4]. Here, we detail the prompts and show an analysis for this metric.

LLM evaluation prompt. Fig. A.1 shows the system, user, and assistant prompts, that we input to GPT-4o-mini [12], to define the sign language translation evaluation task. To calibrate the language model, we include 12 manually annotated in-context examples, displayed in Tab. A.12, with two examples per score from 0 to 5. Each example contains both the score and the reasoning according to our instructions, focusing on key nouns and verbs, while giving less importance to pronouns. This approach makes the metric interpretable, as the LLM outputs detailed reasoning for each score.

LLM evaluation analysis. As discussed in Sec. 4.3, we provide additional analysis and statistics on LLM-based evaluation. A human study was conducted where 5 annotators manually scored a set of 70 translations. Fig. A.2 and Fig. A.3 illustrate the correlation between the average of these human scores, various automatic metrics [11, 13, 15], and our LLM-based evaluation metric. As shown in Fig. A.2, the LLM-based metric exhibits the highest correlation with human judgments. This strong correlation highlights its potential as a useful method for evaluating sign language translations. We further provide qualitative examples for the LLM scores in Fig. A.4.

A.2. Prompt details

As explained in Sec. 3.1 of the main paper, we use five distinct prompts to define the task and to describe each cue. The exact prompts are provided in Tab. A.1. Note that when randomly dropping a cue, we also omit the corresponding prompt.

```

{
  "role": "system",
  "content": "Evaluate how well the candidate sentence aligns with the content and meaning of the reference sentence on a scale of 0 to 5.
    Prioritize key nouns and verbs, while giving less importance to subject, pronouns, adjectives, and adverbs.
    Scoring Rules:
    Score at least 1: If the candidate sentence shares at least one key noun or verb (or their synonyms) with the reference sentence.
    Score at least 3: If the candidate sentence matches most of the key nouns and verbs (or their synonyms) from the reference sentence.
    Score at least 5: If the candidate sentence conveys the same overall meaning as the reference sentence, with only minor differences.
    Note: Do not penalize differences in less important words or variations in sentence structure.
    Focus solely on the essential meaning conveyed by the key nouns and verbs.
    The candidate sentences are sign language translations of a signer signing the reference sentence.
    Try to be liberal in the nouns and verbs you consider."
},
# Example 1
{
  "role": "user",
  "content": "Assign a score from 0 to 5 based on the rules provided.
    Provide your answer in JSON format with keys 'score' (0-5) and 'reason' with a brief explanation.
    DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the JSON string.
    Reference Sentence: It's blind to the genius loci.
    Candidate Sentence: And that's what it means to be dislocated."
},
{
  "role": "assistant",
  "content": "{
    'score': 0,
    'reason': 'No shared key nouns or verbs; the reference mentions 'blind' and 'genius loci', while the candidate mentions 'dislocated'; meanings
    are different.'
  }"
},
# Examples continued...
{
  "role": "user",
  "content": "Assign a score from 0 to 5 based on the rules provided.
    Provide your answer in JSON format with keys 'score' (0-5) and 'reason' with a brief explanation.
    DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the JSON string.
    Reference Sentence:{text_gt}
    Candidate Sentence:{text_pred}"
}

```

Figure A.1. **LLM evaluation prompt:** We provide the input format that we feed to GPT-4o-mini [12] to evaluate the quality of the translated sentence (`text_pred`) by asking the LLM to compare it against the ground truth sentence (`text_gt`). Specifically, we design a system prompt to define the task, and a series of user-assistant prompt pairs to provide input-output examples for calibration. The last user prompt includes the translated sentence to be evaluated. Instructions are repeated at each user prompt. Here, we display only one example (enclosed in between # comment lines to facilitate the reading). In practice, we provide 12 in-context examples, which are listed in Tab. A.12, and the full prompt can be found in the code release.

Type	Prompt
Initial	You are an AI assistant designed to interpret a video of a sign language signing sequence and translate it into English.
Previous sentence	The previous context is the following:
Pseudo-glosses	The following are some possible words present in the sentence:
Background description	Description of the background is:
Visual features	The following are the video tokens:

Table A.1. **Prompt details.** Each cue is accompanied by a specific prompt that explains the task and helps the model differentiate between the various inputs.

A.3. Background description collation

In Fig. A.5, we illustrate two examples to show the process for the background description collation. As explained in Sec. 3.2, in the first step, we extract captions from multiple frames; in the second step, we take the unique words after filtering out stop words.

We further perform several analyses on these background descriptions on the BOBSL training set. First, we measure the similarity between background descriptions and the ground truth translation sentences, and obtain 3.4% IoU, 5.3% precision, and 9.3% recall. We note that the informative signal in the background descriptions may be beyond the exact word overlap. Next, we look at the distribution of parts of speech, revealing 56.1% nouns, 19.4% verbs, 11.8% adjectives, and 7.9% proper nouns. Among the most frequently occurring words, “man”

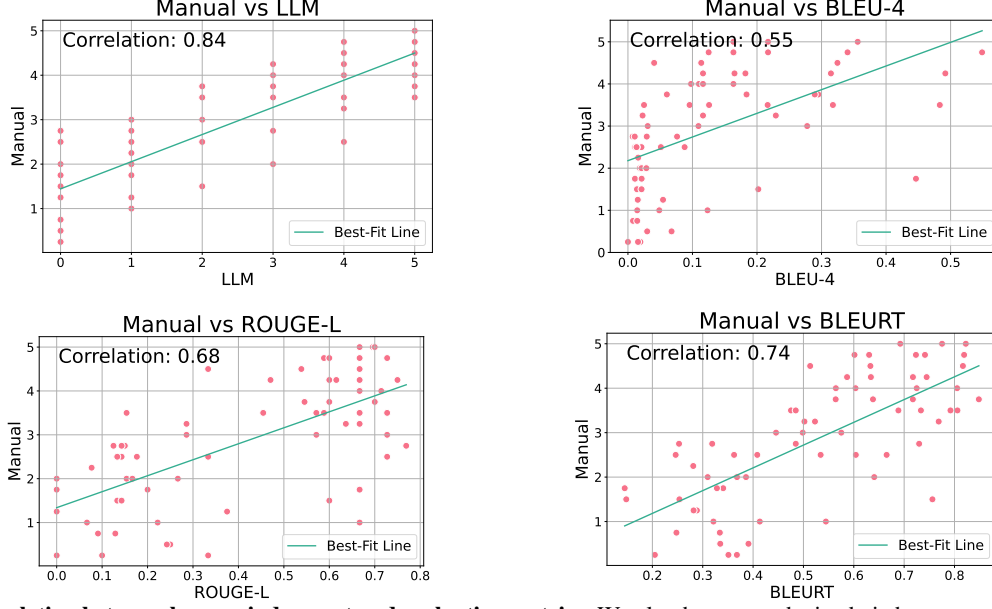


Figure A.2. **Correlation between human judgement and evaluation metrics:** We plot the scores obtained via human evaluation (‘Manual’) against the LLM evaluation scores and the standard captioning metrics (BLEU, ROUGE, and BLEURT). We observe that the LLM score correlates the most with human judgement.

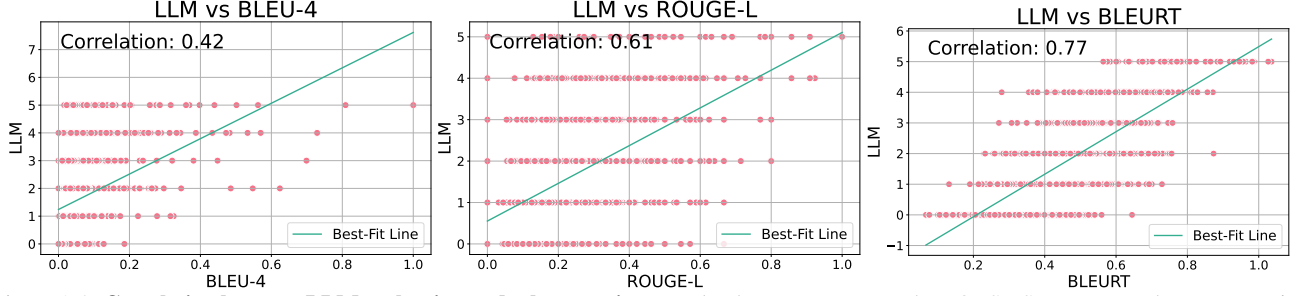


Figure A.3. **Correlation between LLM evaluation and other metrics:** We plot the LLM scores on the BOBSL SENT-VAL and compare against the standard captioning metrics.

was identified as the most common noun, “standing” as the most common verb, and “front” as the most common adjective.

A.4. BOBSL mapping network

The details of the 2-layer MLP used as the mapping network are provided in Tab. A.2. The input to the mapping network consists of Video-Swin features, and its output serves as the input to the LLM decoder. Specifically, in our experiments, the size of the Video-Swin features is 768, while the input size of the LLM decoder (Llama3-8B) is 4,096.

A.5. How2Sign training details

ISLR training details. As mentioned in Sec. 3.2 of the main paper, we fine-tune the Video-Swin model, which is released by [14], with annotations provided by [5]. The training data is automatically annotated from mouthing and dictionary sources, and we set thresholds at 0.75 and 0.5, respectively, to filter the data and enhance its reliability. We train the model on 4 A6000

layer	input sizes	output sizes
fc ₁	$T \times C$	$T \times C'$
gelu	$T \times C'$	$T \times C'$
fc ₂	$T \times C'$	$T \times C'$

Table A.2. **Mapping network architecture for BOBSL training.** We display the 2-layer MLP details, which consists of fully-connected layers and a gelu activation. $C = 768$ represents the number of channels in the Video-Swin features, while $C' = 4,096$ denotes the input size of the LLM decoder. T represents the temporal length of the input feature sequence, which has an average value of 56.

GPUs with a batch size of 24 per GPU, utilising the Adam optimizer [8]. Training is performed in bfloat16 precision. The training spans 30 epochs, including the warmup phase for the first 1 epoch. The learning rate is set to 0.0001 and one cycle cosine learning rate scheduler is adapted.

Visual features. We set the stride (s) to 1 for feature extraction using the Video-Swin model on the How2Sign dataset, as the

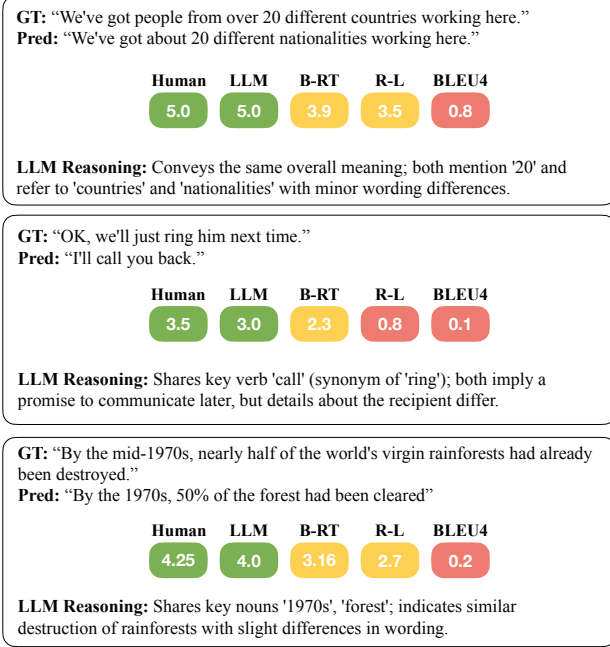


Figure A.4. **Qualitative examples of LLM evaluation:** The scores obtained by LLM strongly correlate with human judgements while simultaneously being able to give detailed descriptions for the reasoning. All scores above are scaled to be out of 5 to make the comparison easier.

layer	kernel	stride	padding	input sizes	output sizes
conv ₁	5	1	2	$T \times C$	$T \times C$
relu ₁	-	-	-	$T \times C$	$T \times C$
maxpool ₁	2	2	-	$T \times C$	$T/2 \times C$
conv ₂	5	1	2	$T/2 \times C$	$T/2 \times C$
relu ₂	-	-	-	$T/2 \times C$	$T/2 \times C$
maxpool ₂	2	2	-	$T/2 \times C$	$T/4 \times C$
fc ₁	-	-	-	$T/4 \times C$	$T/4 \times C'$
gelu	-	-	-	$T/4 \times C'$	$T/4 \times C'$
fc ₂	-	-	-	$T/4 \times C'$	$T/4 \times C'$

Table A.3. **Mapping network for How2Sign training.** We apply 1D CNN on the visual features extracted from the Video-Swin ISLR model. The output of this CNN is then fed into a 2-layer MLP. $C = 768$ represents the number of channels in the Video-Swin features, while $C' = 4,096$ denotes the input size of the LLM decoder. T represents the temporal length of the input feature sequence, which has an average value of 171.

data is smaller and manageable for training. The average number of features is 171, corresponding to a 6.8-second long sequence.

Mapping network on How2Sign. As mentioned in Sec. 3.3 of the main paper, we further provide detailed information about the mapping network for training our model on the How2Sign dataset. Through our experiments, we found that training with only 2-layer MLP was not successful on the How2Sign dataset. Therefore, we add a simple 1D CNN before the MLP layers to compress the long sequences with minimal additional parameters. The 1D CNN is configured with a specific sequence of layers: $\{K5, P2, K5, P2\}$, where K_σ

Model	Size	B4	B-RT	R-L	CIDEr	IoU	LLM
Llama3.2	1B	2.4	39.2	15.8	35.8	13.6	1.05
Llama3.2	3B	3.1	40.0	17.0	41.9	14.6	1.20
Llama3	8B	3.3	40.3	16.9	41.9	14.8	1.20

Table A.4. **LLM decoder variants.** The Llama3-8B model used in the main paper performs overall better than more recent variants of Llama3.2 with less parameters. Note the results are reported on BOBSL SENT-TEST.

denotes a kernel size of σ , and P_σ represents a pooling layer with a kernel size of σ [7]. Details of this mapping network, including input and output sizes, are provided in Tab. A.3.

Target sentence augmentation for How2Sign. We observe overfitting starting from around 5–6 epochs when training with an LLM on the relatively small How2Sign dataset. To further improve the model’s performance, we employ a data augmentation technique that randomly drops 0–20% of the words from the GT sentences.

B. Additional Experiments

We examine performance variations when using different LLM decoders (Appendix B.1), evaluate all possible cue combinations (Appendix B.2), investigate scenarios with missing cues (Appendix B.3), with multiple number of previous sentences (Appendix B.4), and with various background frame sampling rates (Appendix B.5). We also showcase the performance of our ISLR backbone on the HowSign dataset (Appendix B.6), and report the reproduction results of GFSLT [17] and Sign2GPT [16] on the PHOENIX14T dataset (Appendix B.7). Finally, we demonstrate the applicability of our method on PHOENIX14T (Appendix B.8).

B.1. Llama decoder variants

To further analyse the impact of the LLM decoder on performance, we experiment with various Llama variants. Specifically, we compare the Llama3-8B model used in the main paper experiments to more recent and smaller Llama3.2 models: Llama3.2-1B and Llama3.2-3B. As shown in Tab. A.4, Llama3.2-3B demonstrates performance comparable to Llama3-8B. When using the Llama3.2-1B model, we observe a performance drop of 0.7 in the BLEU-4 (B4) score compared to the Llama3-8B model. However, Llama3.2-1B still outperforms all baselines compared in Tab. 3 of the main paper. Note that this experiment is conducted on BOBSL SENT-TEST.

B.2. Combining different cues

We complement Tab. 1 of the main paper by providing results of all possible cue combinations in Tab. A.5. These experiments reveal consistent performance improvements with each added cue, demonstrating that all cues complement each other.



Figure A.5. **Background description collection:** We illustrate, with two examples (left and right blocks), the process of collecting a background description from a signing sentence video. First, we use the BLIP2 image captioner [10] to extract captions for a sequence of background frames in the video. Then, we remove stop words and use the set of unique words to represent the final background description.

Vid	PG	Prev ^{Pred}	BG	B-RT	IoU	LLM
✓				41.0	16.6	1.29
✓	✓			41.8	17.5	1.40
✓		✓		41.5	17.0	1.38
✓			✓	41.9	17.5	1.41
✓	✓	✓		42.5	18.1	1.45
✓	✓		✓	43.2	18.6	1.54
✓		✓	✓	43.1	18.3	1.52
✓	✓	✓	✓	43.5	18.8	1.56

Table A.5. **Combining different cues.** We complement Tab. 1 of the main paper with more combination of cues and report results on BOBSL SENT-VAL. A checkmark ✓ indicates a cue provided during training and testing.

Vid	PG	Prev ^{Pred}	BG	B-RT	IoU	LLM
✓				41.1	17.2	1.31
✓	✓			41.8	18.1	1.41
✓		✓		41.7	17.3	1.39
✓			✓	42.0	17.0	1.42
✓	✓	✓		42.6	17.8	1.49
✓	✓		✓	42.8	18.7	1.52
✓		✓	✓	42.5	17.7	1.52
✓	✓	✓	✓	43.5	18.8	1.56

Table A.6. **Missing cue scenario at test time.** We perform inference using the model trained with all cues. A checkmark ✓ indicates a cue provided during inference, while a blank space denotes a missing cue. Results are reported on BOBSL SENT-VAL.

B.3. Missing cue scenario

As discussed in Sec. 3.3 of the main paper, the *Drop Cue* augmentation enables our model to perform flexible translations even when certain cues are missing during test time. The experimental results are presented in Tab. A.6. Note that, while the previous Tab. A.5 displays the performance of models trained with various cue combinations, Tab. A.6 reports the inference results of the model trained with *all* cues. Notably, our final model achieves results comparable to those of the models trained on specific combinations of cues (i.e. models listed in Tab. 1 of the main paper). This demonstrates that our final model can perform sign language translation with minimal performance degradation when certain cues are unavailable during inference.

B.4. Number of previous sentences

Tab. A.7 reports the results of our best model (Vid+PG+Prev^{Pred}+BG) on SENT-TEST, using 2–3 previous sentences as context during both training and inference. As shown in the table, providing a longer previous context results in only a marginal improvement (+0.1 BLEU-4 and +0.3 ROUGE). While additional context slightly improves B-RT (41.0 vs. 40.3), it comes at the cost of increased computational overhead during inference. To balance performance and efficiency, we use only a single

#prev.	B4	B-RT	R-L	CIDEr	IoU	LLM
1	3.3	40.3	16.9	41.9	14.8	1.20
2	3.3	40.6	17.0	42.8	14.9	1.21
3	3.4	41.0	17.2	43.9	15.1	1.24

Table A.7. **Number of previous sentences (BOBSL SENT-TEST).** We experiment with giving more previous context, and achieve only marginal improvements.

previous sentence for inference.

B.5. Background frame sampling rate

Tab. A.8 reports the results of our best model (Vid+PG+Prev^{Pred}+BG) on SENT-TEST, when varying the sampling rate of the background frames during both training and inference. In the rest of the experiments, we sample a background caption every 1 second. In Tab. A.8, we experiment with reducing this rate by sampling every 2 or 3 seconds. Since we remove repeated words in background captions, we do not experiment with sampling more than 1 frame per second (adjacent frames often depict nearly identical scenes). We obtain (40.3, 40.2, 40.0) BLEURT when sampling every (1, 2, 3) seconds, indicating little effect on performance at lower sampling rates. Thus, sampling ev-

Sampling rate (sec)	B4	B-RT	R-L	CIDEr	IoU	LLM
1	3.3	40.3	16.9	41.9	14.8	1.20
2	3.3	40.2	17.0	41.3	14.8	1.20
3	3.1	40.0	16.9	40.9	14.5	1.19

Table A.8. **Background frame sampling rate (BOBSL SENT-TEST).** We experiment with the background frame sampling rate, by sampling a caption every 1-2-3 seconds, and see little effect on performance when sampling less frames (last row). Note that each sentence lasts on average 4.5 seconds.

Model	Training	Per-instance		Per-class	
		top-1	top-5	top-1	top-5
I3D [5]	BOBSL \rightarrow How2Sign	59.5	78.9	44.5	68.7
Video-Swin (Ours)	How2Sign	63.9	86.0	41.8	69.3
Video-Swin (Ours)	BOBSL \rightarrow How2Sign	77.0	92.8	58.5	82.3

Table A.9. **ISLR performance on How2Sign test set.** Per-instance accuracy is measured over all test instances, while per-class accuracy reflects the average performance across the sign categories in the test set.

ery second avoids missing scene transitions while remaining computationally feasible.

B.6. ISLR performance on How2Sign

The test set provided by [5] is composed of 2,212 manually annotated data. We evaluate both per-instance and per-class accuracy metrics. Per-instance accuracy is calculated across all test instances, while per-class accuracy represents the average performance across the sign categories in the test set. This metric is particularly useful for addressing the unbalanced nature of the datasets, as recommended in [2].

As shown in Tab. A.9, our Video-Swin ISLR model, trained without pre-training on the BOBSL dataset, achieves performance comparable to the I3D ISLR model [5], which is pre-trained on the BOBSL dataset and fine-tuned on the How2Sign dataset. Furthermore, when the Video-Swin ISLR model is initialised with weights pre-trained on the BOBSL dataset, as released by [14], and further fine-tuned on the How2Sign dataset by us, it achieves a 13.1% improvement in per-instance top-1 accuracy and a 16.7% improvement in per-class top-1 accuracy. This underscores the effectiveness and robustness of our framework’s ISLR backbone.

B.7. Reproducing GFSLT and Sign2GPT on PHOENIX14T

As mentioned in Sec. 4.2 of the main paper, we reproduce the performance of the GFSLT and Sign2GPT models on the PHOENIX14T dataset [3, 9]. The results are shown in Tab. A.10. The \dagger symbol denotes the reproduced results, which show comparable performance to the results reported in their original papers across all metrics.

Training on BOBSL. For GFSLT, we observed that using the official codebase leads to gradient divergence during the masked word reconstruction process in text decoding.

Model	B1	B2	B3	B4	R-L
GFSLT [17]	43.71	33.18	26.11	21.44	42.49
GFSLT [17] \dagger	42.02	31.88	25.30	20.76	42.62
Sign2GPT [16]	45.43	32.03	24.23	19.42	45.23
Sign2GPT [16] \dagger	44.14	32.72	25.49	20.82	43.70
Sign2GPT (w/PGP) [16]	49.54	35.96	28.83	22.52	48.90
Sign2GPT (w/PGP) [16] \dagger	46.90	35.72	28.30	23.22	46.28

Table A.10. **Reproducing GFSLT and Sign2GPT on PHOENIX14T.**

\dagger denotes our reproduction results and PGP denotes pseudo-gloss pre-training introduced in [16].

Model	B4	B-RT	R-L	CIDEr	IoU
GFSLT [17]	21.44	-	42.49	-	-
Sign2GPT [16]	19.42	-	45.23	-	-
Sign2GPT (w/PGP) [16]	22.52	-	48.90	-	-
Ours (Vid)	20.58	52.25	41.20	190.40	34.05
Ours (Vid+PG [1])	23.80	52.80	46.11	227.05	38.49

Table A.11. **Evaluation of our model on PHOENIX14T test split.**

Incorporating PG as an additional textual cue improves performance across all evaluation metrics.

To mitigate this issue, we reduced the weight of the word reconstruction loss from 1 to 0.1. For Sign2GPT, as the official codebase only includes the model and hyperparameters, we developed training code using Accelerate [6] framework.

B.8. Evaluation of our model on PHOENIX14T

To evaluate the generalisability of our model, we conduct experiments on the PHOENIX14T dataset [3, 9]. We extract video features using a Video-Swin model trained with pseudo-glosses (PG) obtained from SlowFastSign [1]. As shown in Tab. A.11, when fine-tuning an LLM-based model using only video features, we achieve a BLEU-4 score of 20.58 and a ROUGE-L score of 41.20. By incorporating PG as an additional textual cue, performance improves to a BLEU-4 score of 23.80 and a ROUGE-L score of 46.11. Compared to Sign2GPT [16], which reports a BLEU-4 score of 22.52 and a ROUGE-L score of 48.90, our model demonstrates comparable performance while highlighting the effectiveness of leveraging PG as an additional cue.

Note that no other contextual information (i.e. previous sentence, background) is available for this evaluation. Although PHOENIX14T consists of TV weather broadcasts, the dataset is segmented at the sentence level and lacks context from previous sentences or background information. Moreover, the dataset includes manually annotated glosses and has a restricted vocabulary, leading to performance saturation. This setting therefore does not reflect open-vocabulary tasks, which present greater challenges in real-world scenarios.

1	Reference: It's blind to the genius loci. Candidate: And that's what it means to be dislocated. Score: 0 Reason: No shared key nouns or verbs; the reference mentions 'blind' and 'genius loci', while the candidate mentions 'dislocated'; meanings are different.
2	Reference: She put it by the entrance to the earth so we figure that they like heavy metal or something. Candidate: You've been in a wheelchair for a long time. Score: 0 Reason: No shared key nouns or verbs; the reference talks about 'entrance', 'earth', 'heavy metal', while the candidate mentions 'wheelchair'; meanings are unrelated.
3	Reference: You're coming along to the finale tomorrow? Candidate: I'll have to wait until tomorrow. Score: 1 Reason: Shares the key noun 'tomorrow' but lacks other key content; meanings are somewhat related but differ.
4	Reference: A man's can was open for attack at any point in their life. Candidate: It's not a joke, it's a way of life. Score: 1 Reason: Shares the key noun 'life', but overall meanings are different; reference discusses vulnerability, candidate discusses lifestyle.
5	Reference: Richard called English Nature, who told him that they were natterjack toads, Britain's rarest amphibian. Candidate: Richard, a Nottinghamshire Englishman, is a naturalist. Score: 2 Reason: Shares key nouns 'Richard' and 'English'; candidate omits details about 'natterjack toads' and 'Britain's rarest amphibian'.
6	Reference: Cromwell treated Ireland like the primitive colony he thought it was, moving the Irish off their farms and using the land to pay his soldiers. Candidate: Cromwell was just one of many areas where the IRA set up roadblocks to stop loyalist paramilitaries and farmers from getting through. Score: 2 Reason: Shares key nouns 'Cromwell', 'Ireland', 'farmers'; reference discusses historical actions, candidate discusses modern events; meanings differ.
7	Reference: He sort of guessed it would be 21 maybe 28 days, ended-up being 35. Candidate: He thought it was 21 days. Score: 3 Reason: Matches key nouns 'he', 'thought', '21 days'; candidate conveys a similar time estimation with fewer details.
8	Reference: They get as high as they can off the ground to push the other male down to the floor, and once that male becomes subservient, he slinks off and the dominant male then goes to his female to breed. Candidate: But when one of them is selected, it's the females that can be changed to breed with the new male. Score: 3 Reason: Shares key nouns 'male', 'female', 'breed'; both discuss breeding behaviors, though specifics differ.
9	Reference: It's a lack of understanding on both sides. Candidate: I don't understand that. Score: 3 Reason: Shares key verb 'understand'; both express lack of understanding; candidate is less specific.
10	Reference: OK, we'll just ring him next time. Candidate: I'll call you back. Score: 4 Reason: Shares key verb 'call' (synonym of 'ring'); both involve making a call; minor differences in context.
11	Reference: Really excited. Candidate: I'm so excited. Score: 5 Reason: Conveys the same overall meaning; both express excitement with minor wording differences.
12	Reference: Every day is totally different. Candidate: You know, every day is different. Score: 5 Reason: Conveys the same overall meaning; both state that each day is different with minor phrasing differences.

Table A.12. **LLM evaluation in-context examples:** We display the set of 12 in-context examples provided to GPT-4o-mini [12] to calibrate the LLM evaluation metric. Each reference-candidate pair is provided to the LLM in the user prompt, with the expected output (score and reason) being provided with the assistant role as shown in Fig. A.1.

C. Additional Qualitative Results

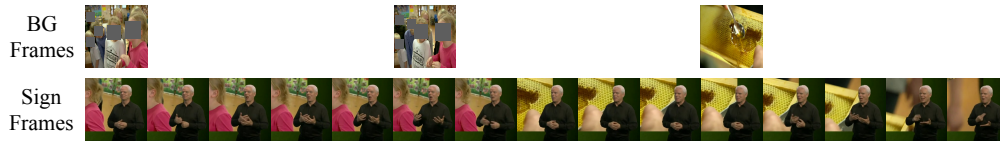
We present additional qualitative results similar to Fig. 3 of the main paper, where we display various inputs, and predictions from our final model compared to baselines. In the first sample of Fig. A.6, the previous sentence indirectly provides information related to location and area, allowing the model to successfully translate the word 'river'. The second sample in Fig. A.6 demonstrates how the background description conveys information about the presence of multiple people on the screen. The third sample in Fig. A.6 demonstrates the ability of the background description to recognise characters on the screen. The first sample in Fig. A.7 shows the model accurately

capturing the object of the sentence from the background description. The second sample in Fig. A.7 demonstrates that the models (Vid+PG, Vid+PG+Prev) make incorrect translations by referring to the 'cliff' word of the pseudo-glosses. However, when all cues are provided, the bias introduced by the pseudo-glosses is resolved. The third sample in Fig. A.7 shows a failure case where the model is biased by background information during the process of pronoun assignment.

Supplementary video. We provide a supplementary video on our project page showcasing several examples comparing against the baselines, along with dynamic sign language videos. The video further includes an example of continuous translation, where consecutive sign language sentences are translated.



Figure A.6. **Qualitative results:** We complement Fig. 3 of the main paper with more examples.



PG : many, lot, well, time, no, good, like, now

Prev : It's a good one.

BG : group, children, standing, classroom, smiling, person, pouring, honey, honeycomb

Vid : There's a lot of it.

Vid+PG : There's a lot of it.

Vid+PG+Prev : A lot of them.

Vid+PG+Prev+BG: There's a lot of **honey** in there.

GT : Got lots of **honey** there



PG : bacon, stick, banana, line, render, and, **cliff**, level, plan, turn, seamount

Prev : What's wrong with these guys?

BG : group, people, standing, dirt, two, women, sitting, front, **tree**, next, **jungle**, man, shirt, necklace

Vid : It's a banana leaf.

Vid+PG : Banana **cliffs**.

Vid+PG+Prev : Banana **cliffs**.

Vid+PG+Prev+BG: Banana **leaves**.

GT : With the banana leaves.



PG : perfect, good, no

Prev : He's looking good.

BG : two, **women**, holding, baby, arms, examining, blue, shirt, woman, koala, blanket

Vid : Perfect.

Vid+PG : Perfect.

Vid+PG+Prev : That's perfect, that's perfect.

Vid+PG+Prev+BG: **She**'s perfect.

GT : That's perfect

Figure A.7. Qualitative results continued

References

- [1] Junseok Ahn, Youngjoon Jang, and Joon Son Chung. Slowfast network for continuous sign language recognition. In *Proc. ICASSP*, 2024. 6
- [2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proc. ECCV*, 2020. 6
- [3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proc. CVPR*, 2018. 6
- [4] David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating image captions with large language models. In *Proc. EMNLP*, 2023. 1
- [5] Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. Sign language video retrieval with free-form textual queries. In *Proc. CVPR*, 2022. 3, 6
- [6] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. <https://github.com/huggingface/accelerate>, 2022. 6
- [7] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proc. CVPR*, 2023. 4
- [8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015. 3
- [9] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 2015. 6
- [10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. ICML*, 2023. 5
- [11] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL*, 2004. 1
- [12] OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2024. 1, 2, 7
- [13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, 2002. 1
- [14] Charles Raude, K R Prajwal, Liliane Momeni, Hannah Bull, Samuel Albanie, Andrew Zisserman, and Gül Varol. A tale of two languages: Large-vocabulary continuous sign language recognition from spoken language supervision. *arXiv*, 2024. 3, 6
- [15] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. BLEURT: Learning robust metrics for text generation. In *Proc. ACL*, 2020. 1
- [16] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. In *Proc. ICLR*, 2024. 4, 6
- [17] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proc. ICCV*, 2023. 4, 6