Supplementary Material for Leveraging the Power of MLLMs for Gloss-Free Sign Language Translation

Jungeun Kim[1*], Hyeongwoo Jeon[2*], Jongseong Bae[1], Ha Young Kim[2†]
[1]Department of Artificial Intelligence, Yonsei University
[2]Graduate School of Information, Yonsei University
{jekim5418, hyeong1204, js.bae, hayoung.kim}@yonsei.ac.kr

This supplementary material provides detailed information on several aspects not included in the main text. Specifically, it includes the implementation details (Sec. A) and the characteristics of the MLLMs employed in the analysis (Sec. B), followed by additional ablation studies (Sec. C). An explanation and examples of failure cases of SL description are provided (Sec. D). Furthermore, we calculate the computational complexity of MLLMs and provide the number of parameters in MMSLT (Sec. E). An outline of the proposed MMSLT algorithm is also included (Sec. F). Additionally, we provide further examples of SL descriptions according to MLLMs and prompts (Sec. G). Finally, additional quantitative results are presented (Sec. H).

A. Training and Inference Details

We use LLaVA-OneVision 7B [8] as our MLLM. To enable efficient and fast SL description generation, we process batches of 8 frames during inference with the MLLM. For the CSL-Daily dataset, input frames are resized to 256×256 before being fed into the MLLM. Furthermore, the maximum output token limit is set to 256, ensuring concise and controlled SL description generation. For description encoder, we use Google's bert-base-cased [5], which distinguishes between uppercase and lowercase letters when embedding sentences. In the modality adapter, the kernel size and stride for the Conv1D are set to 5 and 1, respectively, followed by a BN1D, ReLU, and a MaxPooling1D layer with a kernel size 2 and a stride of 2. For mBART, we employ the official release of mbart-large-50-many-to-many-mmt [14], pre-trained on diverse language datasets. Each SL image is resized to 256×256 and then randomly center-cropped to 224×224. Consistent with a previous study [20], we apply data augmentation at the video level using the VIDAUG [4] library, which includes geometric transformations with a probability of 0.5 during training. For MMLP, we set λ as 0.1 and use the AdamW [13] optimizer with a weight decay of 0.2 and value of betas set to [0.9, 0.98]. We employ a cosine learning rate scheduler that adjusts the learning rate from 1×10^{-4} to a minimum of 1×10^{-8} over 80 epochs with a mini-batch size of 16.

For SLT, we use the AdamW optimizer, weight decay of 1×10^{-3}, betas [0.9, 0.98] and cosine learning rate scheduler with 1×10^{-4} down to a minimum of 1×10^{-8} over 200 epochs with a mini-batch size of 8. We train our network with cross-entropy loss with label smoothing of 0.2. During inference, decoding is performed using the beam search strategy with a length penalty [17] of 1 and a beam size of 8.

All SL description generation and training processes are conducted on 4 NVIDIA RTX 6000 Ada GPUs. For PHOENIX14T dataset, SL description generation takes approximately 2 days, while the entire training process is completed within 24 hours. In the case of the CSL-Daily dataset, SL description generation takes approximately 11 days, and the entire training process is completed in about 6 days.

B. MLLMs Overview

1. LLaVA-Next [9]: A model that demonstrates excellent performance on benchmarks such as VQA within the LLaVA family.
2. InternVL [3]: A model that supports multi-task outputs, including bounding boxes and masks.
3. Qwen2-VL [16]: A model capable of understanding and comparing differences between two images.
4. Pixtral [1]: A model that performs detailed tasks such as diagram understanding and document question answering.
5. LLaVA-OneVision [8]: A model that enhances performance across single-image, multi-image, and video scenarios compared to LLaVA-Next.

[*] These authors contributed equally.

[†] Corresponding author.

C. Additional Ablation Studies

C.1. Quantitative results of video-based MLLMs.

To support our choice of an image-based MLLM, we conducted experiments in which SL descriptions were generated using video-based MLLMs. As shown in the Tab. C.1, this significantly degrades performance, consistent with the qualitative results (Sec. G.1), as video-based MLLMs fail to adequately describe SL elements, hindering accurate understanding of SL semantics.

MLLM	B-1	B-2	B-3	B-4	R
Video-LLaMA 7B [19]	41.29	30.82	24.28	20.09	40.69
Video-LLaVA 7B [11]	47.97	36.78	28.98	23.78	47.20
LLaVA-OneVision 7B [8]	**48.92**	**38.12**	**30.79**	**25.73**	**47.97**

Table C.1. Performance comparison between video-based and image-based MLLMs.

C.2. Type of LLMs

We conduct an ablation study to analyze the performance based on the type of LLMs on the multimodal encoder (ψ_{enc}) and LLM decoder (ψ_{dec}). To this end, we consider a prompt-free multilingual encoder-decoder LLM, as our proposed methodology does not require prompts, and mBART variants, which are commonly used in SLT. As shown in Tab. C.2, mBART50-mmt outperforms mBART25 and mBART50, emphasizing the importance of multilingual capabilities.

Model	B-1	B-2	B-3	B-4	R
NLLB-200-Distilled-600M [15]	39.16	28.61	22.06	18.11	38.30
mBART25 [12]	44.65	34.14	26.87	22.08	44.34
M2M-100-1.2B [6]	45.13	34.68	27.66	22.98	44.78
mBART50 [14]	47.94	36.98	29.51	24.52	47.46
mBART50-mmt [14]	**48.92**	**38.12**	**30.79**	**25.73**	**47.97**

Table C.2. Ablation results of other LLMs and mBART variants.

C.3. Effect of LoRA

We evaluate the effectiveness of LoRA [7] by comparing the performance with and without LoRA applied to the multimodal encoder and the LLM decoder of MMSLT. As shown in Tab. C.3, the performance is lowest when LoRA is not applied to either the encoder or the decoder. Furthermore, applying LoRA to the encoder, which learns alignment with the target spoken sentences, leads to a greater performance improvement than applying it to the decoder, which directly influences the translation output. The best performance is achieved when LoRA is applied to both the encoder and decoder. Consequently, efficiently fine-tuning LLMs for SLT with LoRA contributes to model performance improvement.

ψ_{enc}	ψ_{dec}	B-1	B-2	B-3	B-4	R
-	-	36.97	27.30	21.37	17.53	37.06
-	✓	41.62	30.73	23.95	19.65	39.84
✓	-	47.13	35.34	27.50	22.26	45.95
✓	✓	**48.92**	**38.12**	**30.79**	**25.73**	**47.97**

Table C.3. Ablation results with and without LoRA in ψ_{enc} and ψ_{dec}.

C.4. Temporal Modeling

To effectively learn cross-modal dependencies and joint representations, we designed a framework that concatenates two modalities before temporal modeling, leveraging their complementarity for richer semantic cues. As shown in the Tab. C.4, this approach outperforms modeling the two modalities separately.

Temporal modeling	B-1	B-2	B-3	B-4	R
Individually	48.88	37.76	30.34	25.28	47.85
Jointly	**48.92**	**38.12**	**30.79**	**25.73**	**47.97**

Table C.4. Ablation results of features aligned to the target text.

C.5. Alignment Feature

To identify the optimal features for aligning SL videos and their corresponding descriptions with the target text in the MMLP module, we conduct an ablation study on the outputs of both the modality adapter and the multimodal encoder. As shown in Tab. C.5, the alignment between the multimodal encoder output and the target text embedding yields superior performance. We attribute this improvement to the fact that the feature level remains consistent between the multimodal encoder and the text decoder.

	B-1	B-2	B-3	B-4	R
Modality adapter	47.08	36.26	29.00	24.00	46.52
Multimodal encoder	**48.92**	**38.12**	**30.79**	**25.73**	**47.97**

Table C.5. Ablation results of features aligned to the target text.

C.6. Effect of freezing ψ_{enc}

To verify the necessity of training the ψ_{vis} during the SLT stage, we conducted an experiment in which it was frozen during training. The ψ_{dm} was also kept frozen, as its role is limited to mapping visual features to the SL description space. As shown in the Tab. C.6, the results suggest that fine-tuning the ψ_{vis} is important for capturing task-relevant visual representations.

C.7. Loss function of ψ_{dm}

We perform an ablation study to determine the optimal loss function, $\mathcal{L}_{DM}$, for training the description mapper (ψ_{dm})

Visual encoder	B-1	B-2	B-3	B-4	R
Frozen	43.17	31.99	24.78	19.96	42.34
Trainable	**48.92**	**38.12**	**30.79**	**25.73**	**47.97**

Table C.6. Ablation results with frozen and trainable ψ_{enc}.

to approximate SL descriptions from the visual features of SL videos. To this end, we consider L2 loss and KL divergence. As shown in Tab. C.7, the high BLEU-4 score achieved by L2 loss suggests its effectiveness in enhancing translation, while the high ROUGE score of KL divergence indicates its role in contextual understanding.

	B-1	B-2	B-3	B-4	R
KL divergence	**49.26**	38.02	30.22	24.83	**48.33**
L2 loss	48.92	**38.12**	**30.79**	**25.73**	47.97

Table C.7. Ablation results of the function of L_{DM}.

C.8. MMLP loss coefficient

We conduct a grid search to determine the optimal coefficient (λ) for $\mathcal{L}_{ALIGN}$ in the $\mathcal{L}_{MMLP}$. Fig. C.1 provides a performance comparison of the PHOENIX14T [2] dataset. When the λ values are set to 0.05, 0.1, 0.5, and 0.7, the performance exceeds that of the existing SOTA gloss-free SLT model. However, a significant deterioration in performance is observed when λ is set to 1. Based on these results, the optimal value for λ is 0.1.

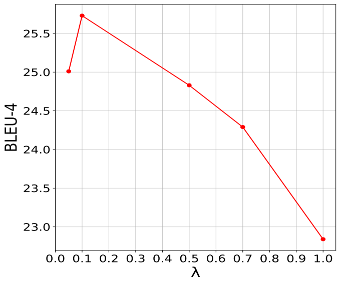
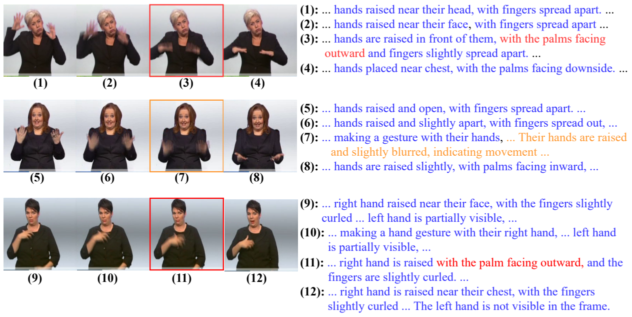


Figure C.1. Performance comparison according to λ, the coefficient of $\mathcal{L}_{ALIGN}$.

D. Failure Cases of SL descriptions

The LLaVA-OneVision 7B model used in this study is not explicitly trained to identify and describe SL elements. Consequently, the MLLM occasionally struggles to accurately capture signer's facial expressions, generating fragmented descriptions of lip shapes, such as "smiling," or misidentifying closed eyes as "open eyes." In addition, we also observed inaccuracies arising in some frames due to blurriness, as shown in Fig. D.1. However, these errors are initially mitigated by combining the visual features of the SL videos with the SL descriptions, which helps compensate for missing SL components. The modality adapter then extracts features by considering the SL description patterns of adjacent frames through short-term modeling. Finally, the multimodal encoder reduces errors by reflecting the continuous patterns of gestures and facial expressions across multiple frames through long-term modeling.



(1): ... hands raised near their head, with fingers spread apart. ...
(2): ... hands raised near their face, with fingers spread apart ...
(3): ... hands are raised in front of them, with the palms facing outward and fingers slightly spread apart. ...
(4): ... hands placed near chest, with the palms facing downside. ...

(5): ... hands raised and open, with fingers spread apart. ...
(6): ... hands raised and slightly apart, with fingers spread out, ...
(7): ... making a gesture with their hands, ... Their hands are raised and slightly blurred, indicating movement ...
(8): ... hands are raised slightly, with palms facing inward, ...

(9): ... right hand raised near their face, with the fingers slightly curled ... left hand is partially visible, ...
(10): ... making a hand gesture with their right hand, ... left hand is partially visible, ...
(11): ... right hand is raised with the palm facing outward, and the fingers are slightly curled. ...
(12): ... right hand is raised near their chest, with the fingers slightly curled ... The left hand is not visible in the frame.

Figure D.1. Example of SL description failure due to blurriness in consecutive frames. The correct parts are shown in blue, the incorrect parts are highlighted in red, and although there is no specific explanation of the SL components, the explanation, which includes the fact that the signer is in movement, is highlighted in yellow.

E. Computation Cost

E.1. Computational Complexity of MLLM

To assess the effectiveness of the description mapper, we calculate the average execution times for generating SL descriptions (Tab. E.1 (1)), translation with actual SL descriptions by integrating the MLLM in the inference loop (Tab. E.1 (2)), and inference using the description mapper (Tab. E.1 (3)). During the measurements, the batch size for videos is set to 1, while the batch size for images is set to 8 for (1) and (2) during MLLM inference. As shown in Tab. E.1, approximating the SL description through the description mapper (3) is 570 times faster than generating an SL description for each input data during inference (2).

		PHOENIX14T [2]	CSL-Daily [21]
(1)	Pre-processing time for generating SL descriptions	107s	215s
(2)	Inference time using actual SL descriptions	255s	429s
(3)	Inference time using the description mapper	**447ms**	**473ms**

Table E.1. Average execution time per sign language video.

E.2. Parameters

Tab. E.2 presents the total number of parameters, the number of trainable parameters, and the percentage of trainable parameters for each component of MMSLT. We exclude the text encoder (ψ_{te}), description encoder (ψ_{de}), and MLLM, which are not trained. While the visual encoder (ψ_{vis}), description mapper (ψ_{dm}), and modality adapter (ψ_{ma}) are fully trainable, the multimodal encoder (ψ_{enc}) and LLM decoder (ψ_{dec}) utilize LoRA, significantly reducing the number of trainable parameters. Specifically, only 0.38% of ψ_{enc} and 1.54% of ψ_{dec} are trainable. This leads to a total of 31.10M trainable parameters, accounting for just 4.85% of the overall model parameters, demonstrating the efficiency of the LoRA approach in reducing computational overhead while improving model performance.

Components	Total (M)	Trainable (M)
Visual encoder (ψ_{vis})	11.18	11.18 (100%)
Description mapper (ψ_{dm})	1.31	1.31 (100%)
Modality adapter (ψ_{ma})	13.90	13.90 (100%)
Multimodal encoder (ψ_{enc})	409.84	1.57 (0.38%)
LLM decoder (ψ_{dec})	204.71	3.14 (1.54%)
Total	640.94	31.10 (4.85%)

Table E.2. Total and trainable parameters, and the ratio of trainable parameters for each component of MMSLT. Values are in millions, with percentages in parentheses.

F. Algorithm of MMSLT

We introduce the algorithms for MMLP and SLT, as shown in Algorithm F.1 and Algorithm F.2.

Algorithm F.1 Multimodal-Language Pre-training

```
Input: Training dataset $\mathcal{D} = \{SV_n, d_n, SL_n\}_{n=1}^{N}$, where $SV$ is a sign language video, $d$ and $SL$ are corresponding SL descriptions and spoken sentence;
Initialize the parameters of $\psi_{vis}(\cdot)$, $\psi_{dm}(\cdot)$, $\psi_{ma}(\cdot)$ and $\psi_{enc}(\cdot)$
Initialize with pre-trained weight and freeze the parameters of $\psi_{de}(\cdot)$, $\psi_{te}(\cdot)$
while not converged do
    for $(SV_i, d_i, SL_i)$ in $\mathcal{D}$ do
        $D_i = \psi_{de}(d_i)$, $V_i = \psi_{vis}(SV_i)$
        Predict SL descriptions $\hat{D}_i = \psi_{dm}(V_i)$
        $\tilde{M}_i = \psi_{enc}(\psi_{ma}(\hat{D}_i \oplus V_i))$, $\tilde{L}_i = \psi_{te}(SL_i)$
        Update $\psi_{vis}(\cdot)$, $\psi_{dm}(\cdot)$ by minimizing the loss $\nabla\mathcal{L}_{DM}(D_i, \hat{D}_i)$+$\nabla\mathcal{L}_{\text{ALIGN}}(\tilde{M}_i, \tilde{L}_i)$
        and
        Update $\psi_{ma}(\cdot)$, and $\psi_{enc}(\cdot)$ by minimizing the loss $\nabla\mathcal{L}_{\text{ALIGN}}(\tilde{M}_i, \tilde{L}_i)$
    end for
end while
Output: Updated $\psi_{vis}(\cdot)$, $\psi_{dm}(\cdot)$, $\psi_{ma}(\cdot)$, and $\psi_{enc}(\cdot)$
```

Algorithm F.2 Sign Language Translation

```
Input: Training dataset $\mathcal{D} = \{SV_n, SL_n\}_{n=1}^{N}$;
Initialize the parameters of $\psi_{vis}(\cdot)$, $\psi_{ma}(\cdot)$, $\psi_{dm}(\cdot)$, and $\psi_{enc}(\cdot)$ with pre-trained in MMLP;
Initialize the parameters of $\psi_{dec}(\cdot)$
Freeze $\psi_{dm}(\cdot)$
while not converged do
    for $(SV_i, SL_i)$ in $\mathcal{D}$ do
        $V_i = \psi_{vis}(SV_i)$
        Predict SL descriptions $\hat{D}_i = \psi_{dm}(V_i)$
        $\tilde{M}_i = \psi_{enc}(\psi_{ma}(\hat{D}_i \oplus V_i))$
        Predict target sentence $\hat{SL}_i = \psi_{dec}(\tilde{M}_i)$
        Update $\psi_{vis}(\cdot)$, $\psi_{ma}(\cdot)$, $\psi_{enc}(\cdot)$, and $\psi_{dec}(\cdot)$ by minimizing the loss $\nabla\mathcal{L}_{SLT}(\hat{SL}_i, SL_i)$
    end for
end while
Output: Updated $\psi_{vis}(\cdot)$, $\psi_{ma}(\cdot)$, $\psi_{enc}(\cdot)$, and $\psi_{dec}(\cdot)$
```

G. Additional Examples of SL Descriptions

In this section, we present and analyze examples of SL descriptions generated by various MLLMs and prompts. In the figures, incorrect parts are highlighted in red, irrelevant information in green, and accurate descriptions in blue.

G.1. Video-based MLLMs

Fig. G.1 illustrates examples of SL descriptions generated using the prompt, "Describe only the motion and gestures of the person in this video, focusing on hands and face," extracted from the PHOENIX14T dataset. The Video-LLaMA tends to repeat sentences with the same content, focuses excessively on the external features of the signer, and exhibits hallucinations due to significant inaccuracies. Additionally, while Video-LLaVA provides more refined responses than Video-LLaMA, it lacks detail regarding the SL components.

G.2. Image-based MLLMs

To evaluate the precision of image-based MLLMs in extracting SL components and the effectiveness of the information they provide, we compare the SL descriptions generated in scenarios where the signers are the same but the actions differ slightly (Fig. G.2) and where the actions are the same but the signers differ (Fig. G.3). The model effectively captures subtle differences in similar. When signers differ, the descriptions of the signer's action remain consistent while omitting unnecessary information.

Fig. G.4 and Fig. G.5 show additional examples with different MLLMs. Among these models, all except LLaVA-OneVision 7B tend to include incorrect information or provide details irrelevant to the meaning of the SL.

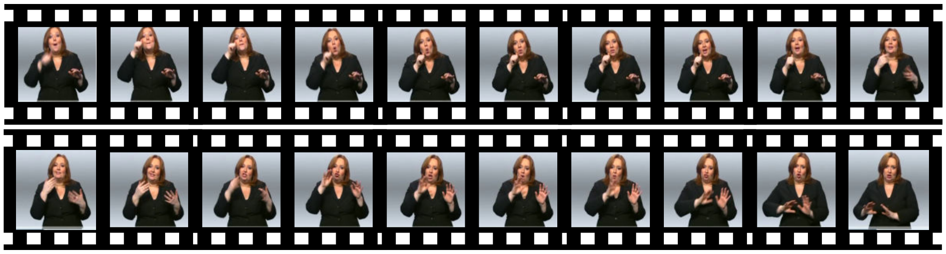
G.3. Prompts

Fig. G.6 and Fig. G.7 show SL descriptions generated by varying prompts for randomly sampled SL images from the PHOENIX14T and CSL-Daily, respectively. In this process, the MLLM consistently uses LLaVA-OneVision 7B [8]. The first, second, and fourth prompts contain significant irrelevant content, while the fifth and sixth prompts exhibit a high prevalence of inaccurate information or simply repeat the examples. In contrast, only the third prompt offers a precise description of the SL component.

H. Additional Qualitative Results

Tab. H.1 and Tab. H.2 present additional translation results randomly selected from the PHOENIX14T and CSL-Daily test sets, respectively. The results are compared against the reference and GFSLT-VLP [20], as in the main text. However, due to a reproducing issue, GFSLT-VLP results are not included for the CSL-Daily dataset. In the tables, correct answers are highlighted in blue, semantically equivalent but lexically different answers are highlighted in green, and incorrect answers are highlighted in red. In all examples, MMSLT generates slightly different words or phrases while preserving the overall meaning.

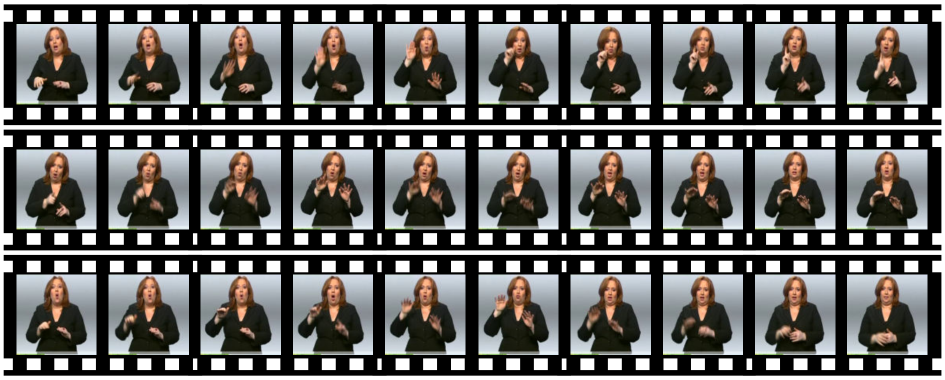
Example: SL descriptions Using Video-based MLLMs



Prompt: Describe only the motion and gestures of the person in the video focus on hands and face.

Video-LLaMA: First, we see a man with red hair pointing his finger at the camera. Next, a woman in a black blazer is shown holding her hand in front of her face. Then, the same woman is talking to someone on a phone while holding her hand up to her mouth. Next, she's seen using her hand to signal to the camera. Then, she's back to talking on the phone. She's talking while holding her hand up to her mouth and holding the phone up to her ear. She's talking to someone on the phone while holding the phone to her ear. The video shows a woman wearing a suit standing with a microphone in her hand. Then, the video ends with the same woman talking to someone on a phone while holding her hand up to her mouth. Throughout the video, the woman is seen holding her hand up to her mouth and gesturing with it. The video is captioned with various descriptions of various objects and actions, such as a white textured background, a woman talking on a phone, a silver necklace, a black and white background, a woman standing on a stage, and more.

Video-LLaVA: The person in the video is seen making various hand gestures, including holding her hand up to her face and pointing her finger at the camera. She also makes a gesture with her hand and mouth, and at one point, she puts her hand on her face.



Prompt: Describe only the motion and gestures of the person in the video focus on hands and face.

Video-LLaMA: First, a woman in a black jacket is seen speaking on a live news show. Next, she is speaking with her hand on her hip. Then, she is wearing a suit while pointing at the screen. A man is also speaking in a suit, with his hand also on his hip. In the following shot, a woman in a suit is seen pointing at the screen. The next woman in a suit is holding her hand in the air. She is also speaking while pointing at the screen. The woman in the black jacket is seen again, this time with her hands on her hips. A woman in a suit is seen pointing at the screen in the next shot. She is also speaking while pointing at the screen. The woman in a suit is then seen with her hand on her hip while pointing at the screen. Finally, a woman with red hair in a suit is seen pointing at the screen while also speaking.

Video-LLaVA: The person in the video is seen making various hand gestures, including holding her hand up to her face and pointing her finger at the camera. She also makes a gesture with her hand and mouth, and at one point, she puts her hand on her face.

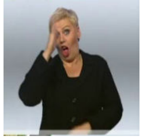
Figure G.1. Additional examples of SL descriptions generated by Video-LLaMA [18] and Video-LLaVA [10] on the PHOENIX14T dataset.

Example: Comparison of SL Descriptions when the Same Signer Performs Similar Signs

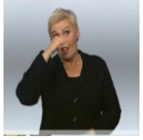
Prompt: Describe only the motion and gestures of the person in the image focus on hands and face.

MLLM:

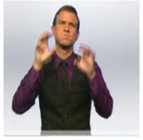
The person in the image has their right hand placed on their forehead, with fingers splayed across the area. Their left hand is partially visible, resting near their waist. The person's mouth is open wide, and their tongue is sticking out slightly. Their facial expression appears to be one of surprise or shock.



The person in the image has their right hand raised to their face, with the index finger touching the nose, suggesting a gesture of contemplation or pointing. Their left hand is partially visible, resting by their side. The person's facial expression appears to be one of surprise or mild shock, with their mouth slightly open and eyes wide.



The person in the image is making a series of hand gestures. Their hands are raised near their face, with fingers extended and slightly bent. The person's facial expression appears to be one of concentration or emphasis, likely correlating with the gestures being made.



The person in the image is making a hand gesture with both hands, with the left hand raised and fingers spread apart, and the right hand partially open with fingers curled inward. The person's mouth is open, and their tongue is sticking out slightly, suggesting they might be speaking or emphasizing a point.

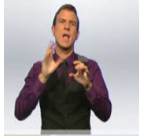


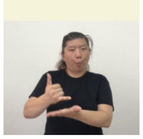
Figure G.2. Additional examples of descriptions generated by an image-based MLLM when the same signer performs similar but slightly different sign gestures on the PHOENIX14T dataset.

Example: Comparison of SL Descriptions when the Different Signer Performs Same Signs

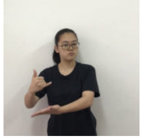
Prompt: Describe only the motion and gestures of the person in the image focus on hands and face.

MLLM:

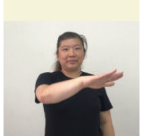
The person in the image is making a hand gesture with their right hand, forming a "hang loose" sign, commonly associated with the "shaka" sign. Their left hand is open with the palm facing forward, fingers spread apart. The person's facial expression appears to be a slight pout or pursed lips, which could indicate a playful or cheeky demeanor.



The person is making a hand gesture with their right hand, forming a "hang loose" sign, and their left hand is open with fingers spread apart. They are also wearing glasses and have a neutral facial expression.



The person in the image has their right hand extended outward with the palm facing forward, suggesting a gesture of offering or presenting something. Their left hand is not visible in the frame. The person's facial expression is neutral, and they are looking directly at the camera.



The person in the image has their right hand raised with the palm facing forward, as if they are about to catch something or are in the middle of a gesture. Their left hand is not visible in the frame. The person is wearing glasses and has a neutral expression on their face.

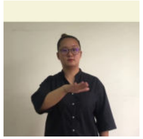
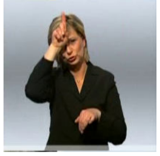


Figure G.3. Additional examples of descriptions generated by image-based MLLMs when the gestures are similar but performed by different signers on the CSL-Daily dataset.

Example: SL descriptions with Various MLLMs

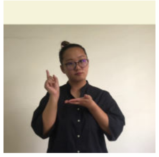


Prompt: Describe only the motion and gestures of the person in the image focus on hands and face.

LLaVA-OneVision 0.5B: The person in the image is making a gesture with their right hand, which appears to be pointing upwards and slightly to the left. Their fingers are spread apart, forming an 'V' shape. The motion suggests they might be emphasizing or highlighting something specific within that gesture.
LLaVA-NeXT 7B: The person in the image is making a gesture with their hand, which appears to be a combination of a peace sign and a pointing gesture. Their thumb and index finger are extended upwards, forming a V shape, while their other fingers are extended downwards, pointing towards the ground. The person's hand is raised slightly above their head, and their mouth is slightly open as if they are speaking or about to speak. The motion suggests a dynamic or expressive action, possibly indicating a point of emphasis or a conclusion to a statement.
LLaVA-OneVision 7B: The person is making a hand gesture with their right hand, pointing upwards with their index finger. Their left hand is slightly raised, with the palm facing outward. The person's facial expression appears to be neutral or slightly serious.
InternVL2 8B: In the image, the individual is using hand gestures to point at their forehead while speaking. The person's left hand is raised, and their index finger is extended upward, tapping the left side of their face. Their right hand appears to be slightly extended, possibly reaching for or pointing at an object or surface. Their expression seems to convey emphasis or make a significant point. The positioning of their hands and the expression on their face suggest that they are likely explaining or emphasizing something important in their conversation. The gestures appear clear and deliberate, indicating the person's desire to make their point effectively.
Qwen2VL 7B: The person in the image is making a gesture with their right hand, holding it up with the index finger extended and the other fingers curled. Their left hand is positioned in front of their chest, with the palm facing outward. The person's face is partially obscured by their right hand, with their eyes looking slightly to the side.
Pixtral 12B : The person in the image is engaged in a dynamic gesture with their hands and face. Their right hand is raised to their forehead, with the fingers spread apart, as if they are shielding their eyes from the sun or emphasizing a point. The left hand is extended outward, palm facing forward, as if signaling a stop or pushing something away. Their face is contorted in an expression of concentration or exasperation, with the mouth slightly open and eyebrows furrowed. The overall posture suggests a moment of intense focus or frustration.

Figure G.4. Additional examples of SL descriptions generated by various image-based MLLMs on the PHOENIX14T dataset.

Example: SL descriptions with Various MLLMs

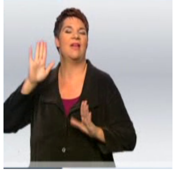


Prompt: Describe only the motion and gestures of the person in the image focus on hands and face.

LLaVA-OneVision 0.5B: The person in the image is making a gesture with both hands, one index finger extended upwards and the other pointing downwards. The fingers are slightly spread apart, indicating a relaxed or open position.
LLaVA-NeXT 7B: In the image, the person is making a gesture with their right hand, which appears to be a "stop" or "pause" sign. Their left hand is resting on their hip, and their face is turned slightly to the side, with a neutral expression. The person is wearing glasses and has a bun hairstyle. The background is plain and does not distract from the subject.
LLaVA-OneVision 7B: The person in the image is making a hand gesture with their right hand, extending the index finger upwards while the other fingers are curled. Their left hand is open with the palm facing upwards, fingers spread apart. The person's facial expression appears neutral, and they are looking directly at the camera.
InternVL2 8B: The person in the image is standing with their hands positioned in front of them at the chest level. They have just made a subtle gesture with their right hand, where they have bent their thumb and index finger slightly outward. Simultaneously, their other fingers are extended, creating a specific hand gesture that might be used for a point or to emphasize a particular message. The person's face is looking directly towards the camera, maintaining a neutral expression. They have styled their hair in a bun and are wearing glasses, along with a dark-colored shirt.
Qwen2VL 7B: The person in the image is making hand gestures. One hand is raised with the index finger pointing upwards, while the other hand is making a circular motion near the chest. The person is also wearing glasses and has their hair tied back.
Pixtral 12B : The person in the image is standing against a plain background. Their hands are positioned in front of them, with the fingers of both hands spread apart and forming a square shape. The person's face is neutral, with their eyes looking directly at the camera. Their glasses rest comfortably on their face, and their hair is neatly tied back. The overall posture is upright and composed.

Figure G.5. Additional examples of SL descriptions generated by various image-based MLLMs on the CSL-Daily dataset.

Example: SL descriptions with Various Types of Prompts



Prompt: Describe about this image.
MLLM: In the image, there is a woman standing against a plain background. She appears to be in mid-gesture, with her right hand raised and palm facing outward as if she is waving or signaling stop. Her left hand is partially raised near her chest, with fingers spread apart. The woman has short hair and is wearing a dark-colored top over a purple garment. Her facial expression suggests that she might be speaking or expressing herself animatedly. The background is a gradient of light colors, providing a neutral backdrop that makes the subject stand out. There are no other objects or people visible in the image.

Prompt: Describe the motion in this image.
MLLM: The image depicts a person who appears to be in the middle of a hand gesture or sign. The individual is wearing a black jacket over a purple top and has short hair. Their right hand is raised with the palm facing outward, suggesting a gesture that could mean "stop" or "no." Meanwhile, their left hand is positioned near their chest, fingers slightly curled as if they are about to clap or make another sign. The background is plain and light-colored, providing no additional context for the gesture being made.Given the nature of the gesture, it's possible that this image is from a video or presentation related to sign language, communication, or perhaps a demonstration of hand movements for a specific purpose. The person's facial expression is not clearly visible due to the angle and lighting, but their body language suggests they are actively communicating or instructing.

Prompt: Describe only the motion and gestures of the person in the image focus onhands and face.
MLLM: The person in the image has their right hand raised with the palm facing outward, as if they are waving or signaling to stop. Their left hand is positioned near their chest, with fingers slightly spread apart, suggesting a gesture of explanation or emphasis. The person's facial expression appears to be one of mild surprise or emphasis, with their mouth slightly open and eyes looking forward.

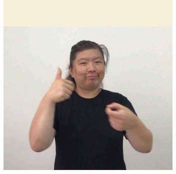
Prompt: This is a frame from a sign language video, and it's important to capture details such as hand gestures and facial expressions accurately. Please identify and describe the specific motion in this frame, including hand position, movement, shape and any relevant facial expressions or emotions shown.
MLLM: The image shows a person performing a sign language gesture. The individual appears to be in the middle of a sequence, with their right hand raised and palm facing outward, which could indicate a stop or halt gesture. Their left hand is positioned near the chest area, fingers slightly curled, possibly indicating a part of a more complex sign or a follow-up gesture after the stop sign. The person's facial expression seems to be one of concentration or emphasis, which is common when signing to ensure clarity and convey the intended message. The background is plain and light-colored, providing no distractions from the sign being performed.

Prompt: Describe the motion in this image.
Example:
Q: Describe the motion in this image.
A: In this image, the right hand is raised to chest level, with fingers slightly bent. The palm is facing the body, while the left hand remains still near the waist, forming a fist. The person's facial expression shows concentration, with slightly furrowed brows and lips pressed together.
MLLM: In this image, the right hand is raised with the palm facing outward and fingers spread apart. The left hand is positioned near the chest, with the fingers slightly curled inward. The facial expression shows a neutral gaze with relaxed eyebrows and lips slightly parted.

Prompt: Describe the motion in this image.
Example:
Q: Describe the motion in this image.
A: Right hand: Raised to chest level, with fingers slightly bent. The palm is facing the body. Left hand: Remains still near the wairs, forming a fist. Face: Concentrated expression with slightly furrowed brows and lips pressed together.
MLLM: Right hand: Fully extended above the head, with fingers spread wide apart. Left hand: Lifted to shoulder height, fingers relaxed and open. Face: Neutral expression with relaxed eyebrows and a slight smile.

Figure G.6. Additional examples of SL descriptions with various types of prompts on the PHOENIX14T dataset.

Example: SL descriptions with Various Types of Prompts



Prompt: Describe about this image.

MLLM: In the image, there is a person standing against a plain background. The individual appears to be smiling and making a thumbs-up gesture with their right hand, which is commonly associated with approval or positivity. They are wearing a black t-shirt and have their hair pulled back. The lighting in the room seems even, without any harsh shadows on the person's face, suggesting an indoor setting with artificial lighting. The background is a solid color, providing no additional context or distractions from the subject.

Prompt: Describe the motion in this image.

MLLM: The image shows a person making a thumbs-up gesture with their right hand, while their left hand is slightly raised and appears to be in motion, possibly gesturing or pointing. The person has a slight smile on their face, suggesting a positive or approving attitude. The background is plain and light-colored, providing a neutral backdrop that focuses attention on the person's actions.

Prompt: Describe only the motion and gestures of the person in the image focus onhands and face.

MLLM: The person is giving a thumbs-up with their right hand and making a fist with their left hand. They have a slight smile on their face.

Prompt: This is a frame from a sign language video, and it's important to capture details such as hand gestures and facial expressions accurately. Please identify and describe the specific motion in this frame, including hand position, movement, shape and any relevant facial expressions or emotions shown.

MLLM: The individual in the image is performing a hand gesture that is commonly associated with the sign for "OK" or "perfect" in American Sign Language (ASL). The right hand is raised with the thumb and index finger touching to form a circle, while the other three fingers are extended upwards. This hand position is known as a "thumbs up" gesture in many cultures, but in ASL, it specifically represents the concept of "OK" or "perfect." The person's facial expression appears to be one of mild amusement or contentment, with a slight smile and eyes looking directly at the camera. There is no additional context provided by the background, which is plain and nondescript, focusing all attention on the hand gesture and the person's expression. Given the simplicity of the background and the directness of the gesture, this frame likely serves an

Prompt: Describe the motion in this image.
Example:
Q: Describe the motion in this image.
A: In this image, the right hand is raised to chest level, with fingers slightly bent. The palm is facing the body, while the left hand remains still near the waist, forming a fist. The person's facial expression shows concentration, with slightly furrowed brows and lips pressed together.

MLLM: A: In this image, the right hand is raised with the thumb extended upwards, indicating a thumbs-up gesture. The left hand is positioned near the chest, with fingers curled inward and slightly bent. The facial expression shows a slight smile with eyes looking directly at the camera.

Prompt: Describe the motion in this image.
Example:
Q: Describe the motion in this image.
A: Right hand: Raised to chest level, with fingers slightly bent. The palm is facing the body. Left hand: Remains still near the wairs, forming a fist. Face: Concentrated expression with slightly furrowed brows and lips pressed together.

MLLM: Right hand: Raised to chest level, with fingers slightly bent. The palm is facing the body. Left hand: Remains still near the wairs, forming a fist. Face: Concentrated expression with slightly furrowed brows and lips pressed together.

Figure G.7. Additional examples of SL descriptions with various types of prompts on the CSL-Daily dataset.

Reference:	in der nordhälfte gibt es zu beginn der nacht noch einzelne heftige schauer und gewitter dann klar dass dort teilweise auf (In the northern half there will be isolated heavy showers and thunderstorms at the beginning of the night, then it will clear up in some places.)
GFSLT-VLP:	in der nacht gibt es im norden noch einzelne schauer oder gewitter die meist abklingen (During the night there will be isolated showers or thunderstorms in the north, which will mostly subside.)
MMSLT:	im norden gibt es zu beginn der nacht noch teilweise kräftige schauer und gewitter später lockern die wolken auf (In the north, there will be some heavy showers and thunderstorms at the beginning of the night; later the clouds will clear up.)
Reference:	deutschland liegt morgen unter hochdruckeinfluss der die wolken weitgehend vertreibt (Germany will be under the influence of high pressure tomorrow which largely drives away the clouds.)
GFSLT-VLP:	deutschland liegt morgen zwischen einem hochdrucksystem im norden und einer hochdruckzone über dem südlichen europa (Tomorrow, Germany will be located between a high pressure system in the north and a high pressure zone over southern Europe.)
MMSLT:	deutschland liegt morgen unter hochdruckeinfluss einer mischung aus wolken und sonne (Germany will be under the influence of high pressure tomorrow with a mixture of clouds and sun.)
Reference:	am tag sechs grad an den alpen und dreizehn grad am oberrhein (During the day six degrees in the Alps and thirteen degrees on the Upper Rhine.)
GFSLT-VLP:	am tag sechs grad im allgäu und dreizehn grad am oberrhein (During the day six degrees in the Allgäu and thirteen degrees on the Upper Rhine.)
MMSLT:	am tag sechs grad in den alpen und dreizehn grad am oberrhein (During the day six degrees in the Alps and thirteen degrees on the Upper Rhine.)
Reference:	heute nacht ist es meist nur locker bewölkt oder klar im nordwesten gibt es auch mal ein paar dichtere wolken (Tonight it will be mostly just partly cloudy or clear, with a few thicker clouds in the northwest.)
GFSLT-VLP:	heute nacht ist es meist nur locker bewölkt oder klar im nordwesten und westen ziehen ein paar wolken (Tonight it will be mostly cloudy or clear with a few clouds in the northwest and west.)
MMSLT:	heute nacht ist es meist nur locker bewölkt oder klar oder nur sternenklar im nordwesten auch mal dichtere wolken (Tonight it will be mostly just partly cloudy or clear, with a few starry clouds in the northwest.)
Reference:	an der nordsee sowie auf den bergen schwere sturmböen (Heavy gusts of wind on the North Sea and in the mountains.)
GFSLT-VLP:	an der nordsee gibt es ab dem nachmittag stürmische böen (There will be stormy gusts on the North Sea from the afternoon onwards.)
MMSLT:	an der nordsee und auf den bergen schwere sturmböen (Heavy gusts of wind on the North Sea and in the mountains.)
Reference:	auch am samstag regnet oder schneit es verbreitet später wird es im westen und norden freundlicher (On Saturday it will rain or snow widely, but later on it will become nicer in the west and north.)
GFSLT-VLP:	auch am samstag regnet oder schneit es verbreitet im norden wird es später freundlicher (On Saturday it will rain or snow widely, but later on it will become nicer in the north.)
MMSLT:	auch am samstag regnet oder schneit es verbreitet später wird es im westen und norden freundlicher (On Saturday it will rain or snow widely, but later on it will become nicer in the west and north.)
Reference:	ich wünsche ihnen einen schönen abend und machen sie es gut (I wish you a nice evening and take care.)
GFSLT-VLP:	jetzt wünsche ich ihnen noch einen schönen abend (Now I wish you a nice evening.)
MMSLT:	ich wünsche ihnen noch einen schönen abend und machen sie es gut (I wish you a nice evening and take care.)
Reference:	ähnliches wetter auch am donnerstag (Similar weather on Thursday.)
GFSLT-VLP:	und nun die wettervorhersage für morgen donnerstag den vierten april (And now the weather forecast for tomorrow, thursday the fourth of april.)
MMSLT:	ähnliches wetter dann am donnerstag (Similar weather on Thursday.)

Table H.1. Visualization of translation results on the PHOENIX14T test set.

Reference:	今天我想吃面条。 (I want to eat noodles today.)
MMSLT:	今天我想吃面条。 (I want to eat noodles today.)
Reference:	他每天回来都很累。 (He always feels very tired when he comes back every day.)
MMSLT:	他每天回来很累。 (He feels very tired when he comes back every day.)
Reference:	天气预报明天下雪，多穿衣服。 (The weather forecast says it will snow tomorrow, so wear more clothes.)
MMSLT:	天气预报明天下雪，多穿衣服。 (The weather forecast says it will snow tomorrow, so wear more clothes.)
Reference:	附近有很多好吃的饭店。 (There are many delicious restaurants nearby.)
MMSLT:	附近有很多好吃的饭店。 (There are many delicious restaurants nearby.)
Reference:	我开车去高铁站接儿子。 (I drove to the high-speed rail station to pick up son.)
MMSLT:	我开车去高铁站接我儿子。 (I drove to the high-speed rail station to pick up my son.)
Reference:	爷爷年龄大，喜欢爬山。 (My grandfather is old and likes climbing mountains.)
MMSLT:	爷爷年龄大，喜欢爬山。 (My grandfather is old and likes climbing mountains.)
Reference:	牛奶可以和咖啡一起喝。 (You can drink milk with coffee.)
MMSLT:	牛奶可以和咖啡一起喝。 (You can drink milk with coffee.)
Reference:	我不去爬山，我有事。 (I'm not going hiking. I'm busy.)
MMSLT:	我不去爬山，我有点事。 (I'm not going hiking. I'm little busy.)
Reference:	这件红色的衣服怎么样？这是新的。 (How about this red dress? It's new.)
MMSLT:	这件红色的衣服怎么样？ (How about this red dress?)
Reference:	那你现在为什么不去？ (So why don't you go now?)
MMSLT:	你今年为什么不去？ (Why don't you go this year?)
Reference:	我给你预订了一个生日蛋糕。 (I ordered a birthday cake for you.)
MMSLT:	我预定了一个生日蛋糕。 (I scheduled a birthday cake.)
Reference:	椅子上有一件衣服，是谁的？ (There is a suit of clothes on the chair. Whose is it?)
MMSLT:	桌子上有一个衣服，是谁的？ (There is a piece of clothes on the table. Whose is it?)
Reference:	小孩子总喜欢模仿大人的动作。 (Children always like to imitate adults' actions.)
MMSLT:	小孩子一直都喜欢看大人的行为。 (Children like to watch adults' behavior all the time.)
Reference:	警察要检查你的身份证。 (The police will check your ID card.)
MMSLT:	警察要检查你的身份证。 (The police will check your ID card.)
Reference:	今天天气有点冷。 (It's a bit cold today.)
MMSLT:	今天天气有点冷。 (It's a bit cold today.)

Table H.2. Visualization of translation results on the CSL-Daily test set.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 1

[2] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018. 3

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[4] DARPA SAIL-ON. Video augmentation techniques for deep learning. https://github.com/darpa-sail-on/videoaug, 2021. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[6] 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, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020. 2

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2

[8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 1, 2, 5

[9] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 1

[10] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6

[11] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. 2

[12] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020. 2

[13] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[14] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning, 2020. 1, 2

[15] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. 2

[16] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. 1

[18] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 6

[19] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023. 2

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881, 2023. 1, 5

[21] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325, 2021. 3