CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation

Ling Zhou[1,2*], Yuhao Chen[1*], Lin Cheng[1], Chengwen Yao[1], Donghui Sun[1†], Xiaoxin Chen[1]

[1]vivo AI Lab
[2]College of Computer Science and Artificial Intelligence, Fudan University, China
{zhoul9090, yhao.chen0617}@gmail.com, donghui.sun@vivo.com

Abstract

Recent sign language translation methods have advanced visual-language alignment, yet autoregressive generation remains vulnerable to unstable policy updates under ambiguous visual evidence. Such instability may cause locally plausible continuations to accumulate into sentence-level semantic deviations, especially when reinforcement learning is used with noisy reward and advantage estimates. Existing PPO-style policy optimization applies confidence-agnostic clipping bounds to all generated tokens, which limits its ability to distinguish between over-amplified confident predictions and low-probability tokens that may still carry useful corrective signals. To address this issue, we propose CAPO-SLT, a reinforcement learning based framework for stable sign language translation generation. CAPO introduces a confidence-aware policy optimization strategy that adapts the upper clipping bound according to old-policy token probability. It applies more conservative updates to high-confidence predictions while preserving update capacity for low-probability positive-advantage tokens. In addition, CAPO caps negative-advantage updates to improve robustness under noisy supervision. This design provides a simple drop-in replacement for standard clipping and stabilizes generation-stage optimization without changing the underlying visual encoder or reward function. Experiments on Chinese and American sign language benchmarks show consistent improvements over pose-based baselines under the same input setting, suggesting that CAPO-SLT provides an effective optimization strategy for stable sign language translation generation.

1 Introduction

Automatic sign language translation (SLT) aims to translate continuous sign language gestures into fluent natural language sentences, thus bridging the communication gap between deaf and hearing communities (Kim et al., 2025; Cooper et al., 2011). This task poses significant challenges, as sign language videos and natural language sentences differ substantially in modality, temporal granularity, and linguistic structure. Recent studies have made notable progress by improving visual-language alignment, with the goal of learning representations that faithfully connect visual observations with textual semantics (Camgoz et al., 2020; Zhou et al., 2023; Li et al., 2025). However, accurate visual-language alignment alone does not fully determine the stability of the final autoregressive translation process, where each generated token can influence subsequent decoding decisions.

Even with well-aligned visual and textual representations, SLT generation remains a sequential decision process. During autoregressive decoding, a locally plausible token can alter the subsequent context and guide later predictions toward a different meaning. This issue is particularly relevant for sign language videos, where visual evidence can be ambiguous due to subtle hand motions, similar gestures, temporal boundary uncertainty, or incomplete non-manual cues (Hwang et al., 2025; Fish and Bowden, 2025). As illustrated in Fig. 1(a), autoregressive decoding (Li et al., 2025) may produce fluent continuations that are locally plausible under the language prior but insufficiently supported by the visual input, allowing local deviations to accumulate and affect sentence-level semantic consistency. This observation suggests that stable SLT generation requires not only cross-modal alignment, but also explicit control over how token-level generation behavior is updated during learning.

Reinforcement learning (RL) provides a natural way to regulate autoregressive generation by optimizing sequence-level translation quality (Jaech et al., 2024). However, applying RL to SLT is non-trivial because the feedback used for policy updates

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 17999–18014
October 24-29, 2026 ©2026 Association for Computational Linguistics

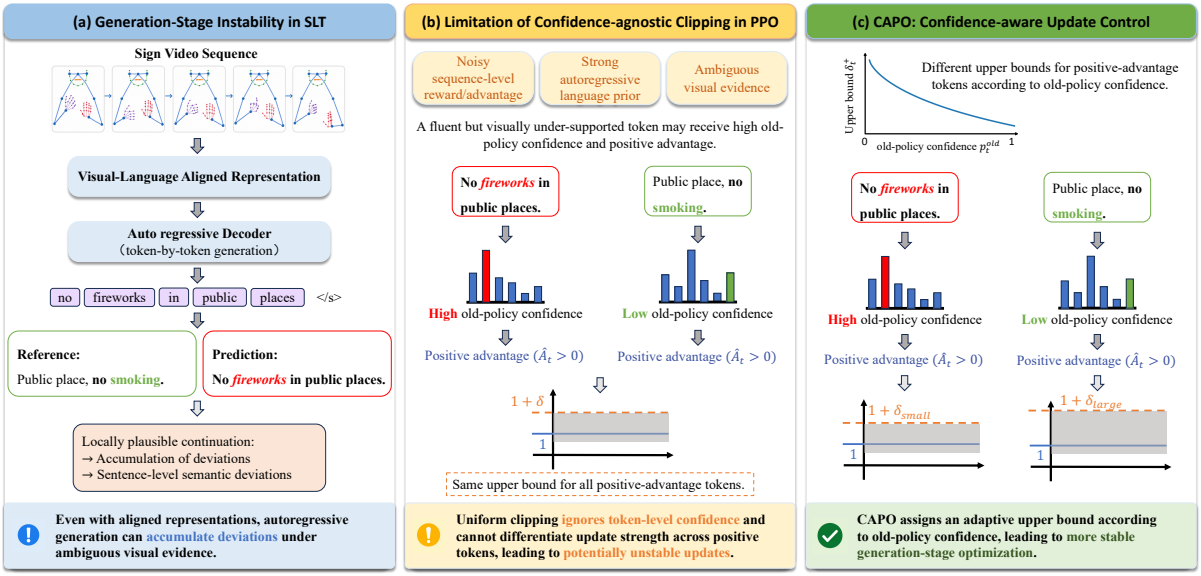


Figure 1: Motivation of CAPO-SLT. (a) Although current SLT methods largely focus on cross-modal alignment, autoregressive decoding may still accumulate locally plausible deviations under ambiguous visual evidence and produce sentence-level semantic deviations. (b) With noisy sequence-level rewards, strong autoregressive language priors, and ambiguous visual evidence, token-level updates can have heterogeneous reliability. Standard PPO-style clipping uses the same upper bound for all positive-advantage tokens, making policy updates insensitive to old-policy confidence and potentially unstable. (c) CAPO assigns confidence-aware upper clipping bounds according to old-policy token probability. It still permits positive updates for high-confidence predictions but limits unnecessary over-amplification, while preserving larger update capacity for low-confidence positive-advantage tokens, enabling more stable generation-stage policy optimization.

is often noisy. Sequence-level rewards only approximate semantic faithfulness, and the resulting advantage estimates must be assigned to individual tokens despite ambiguous visual evidence and strong autoregressive language priors. Standard PPO-style policy optimization mitigates update instability by clipping the policy ratio within a fixed range (Schulman et al., 2017; Shao et al., 2024). While effective as a general trust-region heuristic, this confidence-agnostic clipping rule applies the same upper bound to all positive-advantage tokens. As shown in Fig. 1(b), such a uniform bound does not reflect the heterogeneous token-level update conditions induced by ambiguous visual observations, language-prior driven continuations, and noisy sequence-level reward or advantage estimates (Chen et al., 2023). Consequently, confident predictions may receive unnecessary probability amplification, while low-probability positive-advantage tokens may have limited room for useful updates.

To address this imbalance, we propose CAPO-SLT, a reinforcement learning based framework for stable sign language translation generation. The core of CAPO-SLT is Confidence-Aware Policy Optimization, which uses the old-policy probability of each generated token as an optimization signal to regulate the upper clipping bound for positive-advantage updates. As illustrated in Fig. 1(c), CAPO assigns a tighter upper bound to high-confidence predictions and a more relaxed upper bound to low-probability positive-advantage tokens. Importantly, CAPO does not treat confidence as a direct indicator of semantic correctness, nor does it suppress high-confidence tokens. Since the adaptive upper bound remains above one, high-confidence positive tokens can still be reinforced, but their maximum update ratio is controlled to avoid unnecessary over-amplification. In addition, CAPO caps negative-advantage updates to reduce the effect of noisy negative feedback. In this way, CAPO serves as a drop-in replacement for standard PPO-style clipping and complements cross-modal alignment by stabilizing generation-stage policy optimization.

To validate CAPO-SLT, we conduct experiments on Chinese and American sign language translation benchmarks under matched pose-based input settings. The results show that CAPO-SLT consistently improves translation performance over corresponding baselines. We further analyze policy optimization variants, component contributions, reward configurations, hyperparameter sensitivity,

and training dynamics to examine the robustness and stability of CAPO. Our main contributions can be summarized as follows:

- We revisit SLT from the perspective of generation-stage stability, highlighting that autoregressive translation requires not only cross-modal alignment, but also stable token-level policy updates under noisy sequence-level feedback.
- We propose CAPO-SLT, a reinforcement learning based framework for stable SLT generation. CAPO introduces a confidence-aware policy optimization strategy that adapts the upper clipping bound for positive-advantage tokens according to old-policy token probability, while further capping negative-advantage updates to improve robustness.
- Extensive experiments on both Chinese and American sign language benchmarks demonstrate that CAPO-SLT achieves consistent improvements over corresponding baselines, validating the effectiveness and stability of confidence-aware policy optimization for SLT generation.

2 Related Works

2.1 Sign Language Translation

Sign language translation has been studied under both gloss-based and gloss-free settings. Gloss-based methods use gloss annotations as intermediate linguistic supervision to bridge sign videos and natural language sentences, which helps guide visual representation learning and temporal alignment (Camgoz et al., 2018, 2020; Chen et al., 2022a; Jiao et al., 2024). However, gloss annotations are costly to obtain and may introduce information loss when used as an intermediate representation. Gloss-free methods therefore aim to directly translate sign videos into natural language sentences without relying on gloss supervision (Li et al., 2025). Recent studies in this direction improve SLT through visual-language pretraining, contrastive alignment, spatial-temporal modeling, and large language model based generation (Yin et al., 2023; Zhou et al., 2023; Gong et al., 2024; Kim et al., 2025; Li et al., 2025; Fish and Bowden, 2025). These methods have substantially advanced cross-modal representation learning for SLT by strengthening the correspondence between visual observations and textual semantics. Different from these alignment-oriented studies, our work focuses on the generation-stage policy optimization process. CAPO-SLT complements existing SLT frameworks by regulating token-level policy updates during reinforcement learning fine-tuning, aiming to improve the stability of autoregressive translation generation.

2.2 Reinforcement Learning for Generation

Reinforcement learning has become an important tool for controlling autoregressive generation, especially in language model alignment, instruction following, and reasoning-oriented generation (Schulman et al., 2017; Shao et al., 2024; Guo et al., 2025a). A common choice is PPO-style policy optimization, which stabilizes learning by clipping the policy ratio and preventing excessively large policy updates (Schulman et al., 2017). Recent variants such as GRPO reduce the cost of PPO-style training by replacing value-model based advantage estimation with group-relative rewards, while DAPO further modifies clipping and sampling strategies to improve large-scale reasoning-oriented RL training (Shao et al., 2024; Guo et al., 2025a; Yu et al., 2025). These methods demonstrate the effectiveness of policy optimization for generation, but they are mainly designed for text-only or reasoning-centric scenarios. In SLT, policy updates are conditioned on ambiguous visual evidence, and sequence-level rewards must be assigned to token-level decisions during autoregressive translation (Fish and Bowden, 2025). This creates a different optimization setting, where confidence-agnostic clipping may not sufficiently reflect token-level update conditions. CAPO-SLT is designed for this setting by adapting the positive-advantage upper clipping bound according to old-policy token confidence and by capping negative-advantage updates for more stable SLT generation.

3 CAPO-SLT

3.1 Overview

Fig. 2 illustrates the framework of CAPO-SLT, which consists of two stages. In the first stage, we train a pose-based SLT model with visual-language alignment and supervised translation objectives, which provides a shared initialization for subsequent policy optimization. Given a sign input, the vision encoder and text encoder produce visual and textual representations, which are connected to a LoRA-adapted language model for translation

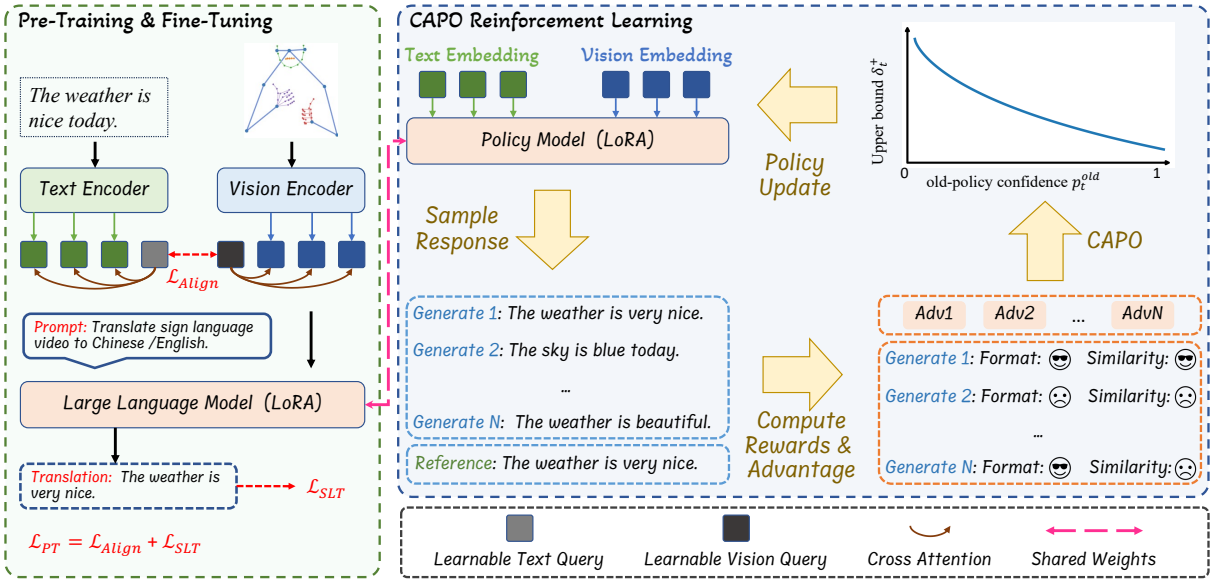


Figure 2: Framework of CAPO-SLT. The framework consists of supervised initialization and CAPO-based reinforcement learning. During pre-training and fine-tuning, the model learns visual-language representations and translation ability through an alignment objective and a supervised SLT objective. During reinforcement learning, the policy model samples multiple candidate translations, receives sequence-level rewards, computes advantages, and updates the policy with CAPO. CAPO adapts the positive-advantage upper clipping bound according to old-policy token confidence and caps negative-advantage updates, providing confidence-aware update control for stable SLT generation.

generation. This stage aims to establish basic cross-modal correspondence and translation ability.

In the second stage, we perform reinforcement learning fine-tuning for autoregressive SLT generation. The policy model samples multiple candidate translations for each sign input, receives sequence-level rewards, and computes advantages for policy updates. CAPO is applied only at this policy optimization stage. Instead of changing the visual encoder or the reward function, CAPO replaces the confidence-agnostic PPO-style clipping rule with confidence-aware update control. This design allows us to study generation-stage optimization while keeping the supervised SLT backbone fixed across policy optimization variants.

3.2 Backbone and Supervised Initialization

Before reinforcement learning, we initialize the policy model with a supervised SLT backbone. Given a sign input, the vision encoder extracts a sequence of visual embeddings, while the text encoder produces textual embeddings from the paired sentence. Then we use learnable visual and textual queries to obtain global representations through attention-based aggregation. These representations are optimized with a visual-language alignment objective, which encourages paired sign and text representations to be close in the shared embedding space.

In addition to alignment learning, the LoRA-adapted language model is trained with a supervised SLT objective under teacher forcing. Let $\mathcal{L}_{align}$ denote the visual-language alignment loss and $\mathcal{L}_{slt}$ denote the supervised translation loss. The supervised initialization objective is

$$\mathcal{L}_{sup} = \mathcal{L}_{align} + \mathcal{L}_{slt}. \tag{1}$$

This stage provides the model with basic cross-modal correspondence and translation ability before reinforcement fine-tuning. Since our focus is policy optimization, the same SLT backbone and supervised initialization are used for all experiments in this paper.

3.3 Confidence-Aware Policy Optimization

After supervised initialization, we further optimize the SLT model with reinforcement learning. Given a sign input x, the policy model π_θ generates a candidate translation $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}$ autoregressively, where each token is generated according to $\pi_\theta(\hat{y}_t|x, \hat{y}_{<t})$. Following PPO-style policy optimization (Schulman et al., 2017), we keep a frozen old policy $\pi_{\theta_{old}}$ during each update step and define the token-level policy ratio as

$$r_t(\theta) = \frac{\pi_\theta(\hat{y}_t \mid x, \hat{y}_{<t})}{\pi_{\theta_{\text{old}}}(\hat{y}_t \mid x, \hat{y}_{<t})}. \tag{2}$$

The sequence-level reward is converted into an advantage estimate $\hat{A}_t$ for valid response tokens. Standard PPO-style optimization clips the policy ratio within a fixed interval, which can be written as

$$\ell_t^{\text{ppo}}(\theta) = -\min\Big(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\delta, 1+\delta\right)\hat{A}_t\Big). \tag{3}$$

This clipping rule provides a simple trust-region constraint, but its upper bound is shared by all positive-advantage tokens. Therefore, it does not distinguish between tokens that already have high old-policy probability and tokens that are assigned low probability but receive positive advantage under the sampled response. In visually grounded SLT generation, where the reward and advantage estimates may be affected by ambiguous visual evidence and language-prior driven continuations, this confidence-agnostic upper bound can lead to less controlled positive updates.

CAPO addresses this issue by adapting the upper clipping bound according to the old-policy probability of each generated token. We define the old-policy token confidence as

$$p_t^{\text{old}} = \pi_{\theta_{\text{old}}}(\hat{y}_t \mid x, \hat{y}_{<t}). \tag{4}$$

We emphasize that p_t^{old} is used only as an optimization signal for controlling update magnitude. It is not treated as a direct indicator of semantic correctness. For positive-advantage tokens, CAPO assigns a token-specific upper clipping range

$$\delta_t^+ = \text{clip}\left(\frac{\delta_{\max}}{1+\alpha(p_t^{\text{old}})^\gamma}, \delta_0, \delta_{\max}\right), \quad \alpha = \frac{\delta_{\max}}{\delta_0} - 1. \tag{5}$$

Here, δ_0 is the base upper clipping range, $\delta_{\max}$ is the maximum relaxed range. Following prior works (Guo et al., 2025a; Yu et al., 2025), we set $\delta_0 = 0.2$ and $\delta_{\max} = 0.3$ to keep the adaptive clipping range within a stable and commonly used scale. γ controls the shape of the confidence-aware schedule. This formulation makes the upper clipping range decrease as old-policy confidence increases. When p_t^{old} is high, δ_t^+ approaches the conservative range δ_0, limiting unnecessary probability amplification for predictions that are already favored by the old policy. When p_t^{old} is low, δ_t^+ moves toward $\delta_{\max}$, preserving larger update capacity for low-probability tokens with positive advantage. Since the effective upper bound is $1+\delta_t^+$, which remains larger than one, high-confidence positive-advantage tokens can still be reinforced. CAPO only controls their maximum update ratio rather than suppressing them.

Using the adaptive upper range, the clipped policy ratio is

$$\tilde{r}_t(\theta) = \text{clip}\left(r_t(\theta), 1-\delta^-, 1+\delta_t^+\right), \tag{6}$$

where δ^- is a fixed lower clipping range. The CAPO clipped loss for positive-advantage update control is then defined as

$$\ell_t^{\text{clip}}(\theta) = -\min\left(r_t(\theta)\hat{A}_t,\ \tilde{r}_t(\theta)\hat{A}_t\right). \tag{7}$$

For tokens with non-positive advantages, we keep the adaptive upper range at the base value and use the fixed lower clipping range. This keeps the confidence-aware relaxation focused on positive-advantage updates, which are responsible for increasing token probabilities.

To improve robustness under noisy negative feedback, CAPO further caps the loss magnitude for negative-advantage tokens. Specifically, for $\hat{A}_t < 0$, we define a constant cap

$$\ell_t^{\text{cap}}(\theta) = -c\hat{A}_t, \tag{8}$$

where $c > 0$ controls the maximum effective loss scale for negative-advantage updates. The final token-level loss is

$$\ell_t(\theta) = \begin{cases} \ell_t^{\text{clip}}(\theta), & \hat{A}_t \geq 0, \\ \min\left(\ell_t^{\text{clip}}(\theta), \ell_t^{\text{cap}}(\theta)\right), & \hat{A}_t < 0. \end{cases} \tag{9}$$

This cap does not reverse the update direction. It only prevents negative-advantage tokens from producing excessively large loss values when the sequence-level reward assigns noisy penalties to individual tokens.

Finally, given a response mask $m_t \in \{0, 1\}$ that indicates valid generated tokens, the CAPO policy optimization loss is computed as

$$\mathcal{L}_{\text{CAPO}}(\theta) = \frac{\sum_{t=1}^{T} m_t \ell_t(\theta)}{\sum_{t=1}^{T} m_t + \epsilon}, \tag{10}$$

where ϵ is a small constant for numerical stability. CAPO is therefore a drop-in replacement for the

Table 1: SLT performance on the CSL-Daily dataset. The best results are highlighted in Red, and the second-best results in Blue. 'Pose' and 'RGB' indicate the input modalities.

Methods	Year	Modality		Dev Set			Test Set		
		Pose	RGB	B-1	B-4	R-L	B-1	B-4	R-L
Gloss-Based Methods									
SLRT (Camgoz et al., 2020) (Gloss-Based)	CVPR20	-	✓	37.47	11.88	37.96	37.38	11.79	36.74
TS-SLT (Chen et al., 2022b)	Neurips22	✓	✓	55.21	25.76	55.10	55.44	25.79	55.72
MMTLB (Chen et al., 2022a)	CVPR22	-	✓	53.81	24.42	53.38	53.31	23.92	53.25
Gloss-Free Methods									
SLRT (Camgoz et al., 2020) (Gloss-Free)	CVPR20	-	✓	21.03	4.04	20.51	20.00	3.03	19.67
GASLT (Yin et al., 2023)	CVPR23	-	✓	-	-	-	19.90	4.07	20.35
GFSLT-VLP (Zhou et al., 2023)	ICCV23	-	✓	39.20	11.07	36.70	39.37	11.00	36.44
SignLLM (Gong et al., 2024)	CVPR24	-	✓	42.45	12.23	39.18	39.55	15.75	39.91
MSLU (Zhou et al., 2025)	TPAMI25	✓	-	33.28	10.27	33.13	33.97	11.42	33.80
SpaMO (Hwang et al., 2025)	NAACL25	-	✓	-	-	-	48.90	20.55	47.46
C^2RL (Chen et al., 2025)	TCSVT25	-	✓	-	-	-	49.32	21.61	48.21
MMSLT (Kim et al., 2025)	ICCV25	-	✓	50.05	20.51	48.53	49.87	21.11	48.92
PGG-SLT (Guo et al., 2025b)	Neurips25	-	✓	53.58	24.05	52.34	53.29	23.70	52.88
Uni-Sign (Li et al., 2025)	ICLR25	✓	-	53.24	25.27	54.34	53.86	25.61	54.92
Uni-Sign (Li et al., 2025)	ICLR25	✓	✓	55.30	26.25	56.03	55.08	26.36	56.51
Geo-Sign (Fish and Bowden, 2025)	Neurips25	✓	-	55.57	27.05	57.27	55.89	27.42	57.95
CAPO-SLT (Ours)	EMNLP26	✓	-	58.83	28.55	58.89	58.82	28.68	58.59

Table 2: SLT performance on the How2Sign dataset. The best results are highlighted in Red, and the second-best results in Blue. 'Pose' and 'RGB' indicate the input modalities.

Methods	Modality		Test Set		
	Pose	RGB	B-1	B-4	R-L
GloFE-VN (Lin et al., 2023)	✓	-	14.9	2.2	12.6
MSLU (Zhou et al., 2025)	✓	-	20.1	2.4	17.2
SLT-IV (Tarrés et al., 2023)	-	✓	34.0	8.0	-
C^2RL (Chen et al., 2025)	-	✓	29.1	9.4	27.0
FLa-LLM (Chen et al., 2024)	-	✓	29.8	9.7	27.8
SSVP-SLT (Rust et al., 2024)	-	✓	43.2	15.5	38.4
PGG-SLT (Guo et al., 2025b)	-	✓	40.8	13.7	32.9
Uni-Sign (Li et al., 2025)	✓	-	40.4	14.5	34.3
Uni-Sign (Li et al., 2025)	✓	✓	40.2	14.9	36.0
Geo-Sign (Fish and Bowden, 2025)	✓	-	40.8	15.1	35.4
CAPO-SLT	✓	-	41.4	15.2	34.9

Table 3: ISLR results on the WLASL2000. The best results are highlighted in Red, and the second-best results in Blue. 'Pose' and 'RGB' indicate the input modalities. † denotes results reproduced by (Hu et al., 2021a).

Methods	Modality		Test Set	
	Pose	RGB	P-I	P-C
ST-GCN† (Yan et al., 2018)	✓	-	34.40	32.53
SignBERT (Hu et al., 2021a)	✓	-	39.40	36.74
HMA (Hu et al., 2021b)	-	✓	37.91	35.90
BEST (Zhao et al., 2023)	✓	-	46.25	43.52
SignBERT+ (Hu et al., 2023)	✓	-	48.85	46.37
MSLU (Zhou et al., 2025)	✓	-	56.29	53.29
NLA-SLR (Zuo et al., 2023)	✓	✓	61.05	58.05
Sign-Rep (Wong et al., 2025)	✓	✓	61.05	58.89
Uni-Sign (Li et al., 2025)	✓	-	63.13	60.90
Uni-Sign (Li et al., 2025)	✓	✓	63.52	61.32
Geo-Sign (Fish and Bowden, 2025)	✓	-	63.64	61.89
CAPO-SLT	✓	-	63.77	61.91

fixed clipping rule in PPO-style optimization. It preserves the proximal update structure while making positive-advantage upper bounds confidence-aware and bounding noisy negative-advantage updates, leading to more stable generation-stage policy optimization for SLT.

3.4 Reward Design for RL Fine-Tuning

During RL fine-tuning, each sampled translation is evaluated by a sequence-level reward. SLT generation requires both surface-form fidelity and semantic consistency, so we use a lightweight combination of format-oriented and semantic similarity signals. Specifically, the format reward $R_{\text{fmt}}(\hat{y}, y)$ is computed as the average of B-1 (Papineni et al., 2002), B-4 (Papineni et al., 2002), and R-L (Lin, 2004) between the generated translation $\hat{y}$ and the reference y. This term encourages lexical overlap and phrase-level fidelity. Since n-gram based metrics can be sensitive to surface variations, we further use BERTScore (Zhang et al., 2019) as the semantic reward $R_{\text{sem}}(\hat{y}, y)$, which provides a complementary signal for contextual semantic similarity. The final reward is defined as

$$R(\hat{y}, y) = \lambda R_{\text{fmt}}(\hat{y}, y) + (1-\lambda) R_{\text{sem}}(\hat{y}, y), \quad (11)$$

where λ controls the balance between surface-form matching and semantic similarity.

Following GRPO (Shao et al., 2024), we compute the advantage of each candidate SLT translation by normalizing its reward within the sampled response group.

Table 4: Comparison with policy optimization variants on the CSL-Daily test set. All RL variants are conducted under the same settings.

Method	Clipping strategy	B-1	B-4	R-L
SFT	None	49.18	20.04	49.40
GRPO	Regular	58.47	27.90	57.88
DAPO	Larger	58.52	28.18	57.92
CAPO	Confidence-aware	**58.82**	**28.68**	**58.59**

Table 5: Ablation of the negative-advantage cap on CSL-Daily test set. Both variants use the same confidence-aware positive upper bound and differ only in whether the negative cap is applied.

Negative Cap	B-1	B-4	R-L
w/o	58.48	28.03	57.80
w/	**58.82**	**28.68**	**58.59**

Table 6: Effect of the reward weight λ on CSL-Daily test set. λ controls the balance between the format reward and the semantic reward.

Metric	0.5	0.6	0.7	0.8	0.9
B-1	58.28	58.44	58.49	**58.82**	58.81
B-4	27.85	28.09	27.94	**28.68**	28.17
R-L	58.17	58.04	57.90	**58.59**	58.13

Table 7: Sensitivity analysis of the scheduling parameter γ on the CSL-Daily test set.

Metric	0.3	0.5	0.7	1.0	1.5
B-1	58.14	58.42	**58.82**	58.57	58.69
B-4	27.70	28.19	**28.68**	28.21	28.08
R-L	57.80	58.29	**58.59**	58.01	57.90

4 Experiments

4.1 Experimental Settings

Datasets. We conduct evaluations of CAPO-SLT on both Chinese Sign Language (CSL) and American Sign Language (ASL) benchmarks. For CSL, experiments are carried out on the CSL-Daily dataset (Zhou et al., 2021). For ASL, we assess translation performance on the How2Sign dataset (Duarte et al., 2021), while isolated sign language recognition (ISLR) is additionally evaluated using the WLASL2000 dataset (Li et al., 2020). More details are provided in Appendix B.1.

Evaluation Metrics. Following previous works (Li et al., 2025; Fish and Bowden, 2025), we report BLEU (B-1, B-4) (Papineni et al., 2002) and ROUGE-L (R-L) (Lin, 2004) scores for SLT, and Per-Class (P-C) & Per-Instance (P-I) accuracy for ISLR.

Implementation Details. Following prior work (Li et al., 2025; Fish and Bowden, 2025), we employ the CSL-News (Li et al., 2025) and YouTube-ASL (Uthus et al., 2023) datasets for the pretraining stage on CSL and ASL, respectively. More details can be found in Appendix B.2

4.2 Comparison with State-of-the-Art

Results on CSL. Table 1 reports the comparison on the CSL-Daily dataset. CAPO-SLT achieves the best performance on both the dev and test sets across all evaluation metrics. On the test set, CAPO-SLT obtains 58.82 B-1, 28.68 B-4, and 58.59 R-L, outperforming the strongest prior method Geo-Sign by 2.93, 1.26, and 0.64 points, respectively. A similar trend is observed on the dev set, where CAPO-SLT improves over Geo-Sign by 3.26 B-1, 1.50 B-4, and 1.62 R-L. Notably, CAPO-SLT uses only pose inputs, yet it surpasses both RGB-based methods and multi-modal methods that rely on pose and RGB inputs. Compared with Uni-Sign under the same pose-only modality, CAPO-SLT brings substantial gains of 4.96 B-1, 3.07 B-4, and 3.67 R-L on the test set. These results demonstrate the effectiveness of CAPO-SLT on Chinese SLT and highlight its strong competitiveness under the pose-only setting.

Results on ASL. Tables 2 and 3 report the results on American sign language benchmarks. On How2Sign, CAPO-SLT achieves 41.4 B-1, 15.2 B-4, and 34.9 R-L using only pose inputs. Our method improves over all pose-based methods on B-1 and B-4, while its R-L score is slightly lower than Geo-Sign. It also performs marginally below the best RGB-based method SSVP-SLT (Rust et al., 2024), which benefits from RGB inputs and a larger-scale visual pre-training setting. Compared with pose-only Uni-Sign, CAPO-SLT brings gains of 1.0 B-1, 0.7 B-4, and 0.6 R-L. On WLASL2000, CAPO-SLT achieves the best performance with 63.77 P-I and 61.91 P-C, outperforming Geo-Sign while using the same pose-only modality. These results indicate that CAPO-SLT generalizes to ASL benchmarks and provides consistent gains under pose-based settings.

4.3 Ablation Studies

Comparison with Policy Optimization Variants. Table 4 compares CAPO with representative policy optimization variants on the CSL-Daily test set. To isolate the effect of the update rule, all RL variants

Table 8: Qualitative examples on CSL-Daily. Words in green indicate accurate translations, words in blue indicate semantically close expressions, and words in red indicate incorrect or visually unsupported content.

Case 1	
Reference	我 分别 介绍 这几位同事。 I introduced these colleagues separately.
w/o CAPO	我 是一个由几个人组成的团队。 I am a team made up of several people.
CAPO	我 分别 介绍了 几位同事。 I introduced several colleagues separately.
Case 2	
Reference	天黑了， 我 害怕。 It is dark, and I am afraid.
w/o CAPO	天黑了， 我 睡着了。 It is dark, and I fell asleep.
CAPO	天黑了， 我 害怕。 It is dark, and I am afraid.
Case 3	
Reference	这部电影 的女演员很漂亮。 The actress in this movie is very beautiful.
w/o CAPO	这部电影 讲的是女演员的爱情故事。 This movie is about the actress's love story.
CAPO	这部电影 的女演员很漂亮。 The actress in this movie is very beautiful.

are initialized from the same supervised checkpoint and use the same reward design and training recipe. GRPO (Shao et al., 2024) with a regular clipping bound substantially improves over SFT, indicating that sequence-level RL fine-tuning is beneficial for SLT generation. DAPO (Yu et al., 2025) further enlarges the upper bound and gives slightly better B-4, but the overall improvement remains limited. CAPO achieves the best results across all metrics, with 58.82 B-1, 28.68 B-4, and 58.59 R-L. These results suggest that, under the same RL fine-tuning setting, confidence-aware upper-bound control provides a more effective update rule for SLT generation than simply using a uniform regular or enlarged clipping bound.

Effect of Negative-advantage Cap. Table 5 isolates the effect of the negative-advantage cap in CAPO. Both variants use the same confidence-aware positive upper bound, and the only difference is whether the cap is applied to negative-advantage updates. Removing the cap leads to lower performance across all metrics, while the full CAPO achieves 58.82 B-1, 28.68 B-4, and 58.59 R-L. This indicates that bounding negative-advantage updates provides additional robustness under noisy feedback and complements the confidence-aware positive update control.

Comparison of Different Clipping Schedules To investigate whether the effectiveness of CAPO stems from the specific confidence-to-bound mapping rather than simply the clipping range, we compare the proposed inverse schedule with a linear confidence-aware schedule on CSL-Daily. Both schedules use the same clipping range ($[\delta_0, \delta_{\max}] = [0.2, 0.3]$) and training configuration. As shown in Table 9, the inverse schedule achieves 58.82/28.68/58.59 on B-1/B-4/R-L, outperforming the linear schedule with 58.66/27.95/57.99. Particularly, the inverse schedule yields improvements of 0.73 and 0.60 points on B-4 and R-L, respectively. Since the overall clipping range is controlled, these results demonstrate that the allocation schedule itself plays an important role. Compared with the linear mapping, the inverse schedule provides a more selective relaxation by allocating larger update capacity primarily to low-probability positive-advantage tokens, while remaining more conservative for medium- and high-confidence tokens. This selective allocation gives low-probability positive-advantage tokens more room for updates while keeping medium- and high-confidence tokens more conservative, which is consistent with the design motivation of CAPO.

Table 9: Ablation of the confidence-to-bound mapping on the CSL-Daily test set. Both schedules use the same clipping range and training configuration.

Schedule	B-1	B-4	R-L
Linear	58.66	27.95	57.99
Inverse	**58.82**	**28.68**	**58.59**

Table 10: Comparison of alternative confidence signals for CAPO on the CSL-Daily test set. TM, TE, and OP denote token margin, token entropy, and old-policy probability, respectively. Results are reported as mean ± standard deviation over three random seeds (42, 43, and 44).

Signal	B-1	B-4	R-L
TM	$58.23_{\pm 0.01}$	$27.75_{\pm 0.00}$	$57.84_{\pm 0.00}$
TE	$58.71_{\pm 0.00}$	$28.03_{\pm 0.00}$	$58.06_{\pm 0.00}$
OP	$\mathbf{58.85}_{\pm 0.01}$	$\mathbf{28.65}_{\pm 0.00}$	$\mathbf{58.63}_{\pm 0.01}$

Comparison of Alternative Confidence Signals We further investigate the choice of confidence signal for modulating the clipping bound in CAPO. In addition to the old-policy probability used in CAPO, we consider two alternatives derived from the policy distribution, namely token margin and token entropy. All variants are trained under the same RL configuration and evaluated on the CSL-

Daily test set using three random seeds (42, 43, and 44), with the results summarized in Table 10. Old-policy probability consistently achieves the best performance across all metrics. Compared with token entropy, it improves B-1, B-4, and R-L by 0.14, 0.62, and 0.57 points, respectively. Token margin yields lower performance, indicating that the separation between the top candidate probabilities is less effective for regulating the update of the sampled token. The superior performance of old-policy probability is also consistent with its role in policy optimization. Rather than estimating semantic correctness, it directly reflects how strongly the old policy favors the sampled token and thus provides a natural token-level signal for controlling the update magnitude.

Effect of Reward Weight. Table 6 studies the influence of the reward mixing weight λ on CSL-Daily. CAPO-SLT performs stably across a range of λ values, suggesting that the method is not overly sensitive to the exact balance between the format reward and the semantic reward. Among these settings, $\lambda = 0.8$ achieves the best overall performance, with 58.82 B-1, 28.68 B-4, and 58.59 R-L. When λ is reduced, the model receives relatively weaker surface-form guidance, which slightly lowers BLEU-based scores. When λ is further increased to 0.9, the performance also drops slightly, indicating that placing too much emphasis on overlap-based rewards may not provide the best balance. Based on these results, we set $\lambda = 0.8$.

Sensitivity to scheduling parameter. Table 7 reports the sensitivity of CAPO to the scheduling parameter γ, which controls the curvature of the confidence-aware upper-bound schedule. CAPO maintains competitive performance across different γ values, showing that its effectiveness is not tied to a single narrow setting. Among the tested values, $\gamma = 0.7$ achieves the best overall results, with 58.82 B-1, 28.68 B-4, and 58.59 R-L. Smaller values such as $\gamma = 0.3$ lead to weaker performance, while larger values also show mild degradation. We therefore set $\gamma = 0.7$ in the main experiments.

4.4 Qualitative Analysis

Table 8 shows representative examples from CSL-Daily. Compared with the model without CAPO, CAPO produces translations that better preserve key semantic units such as predicates, entities, and event relations. The baseline outputs are often fluent, but may contain meaning shifts or visually unsupported details. In contrast, CAPO reduces such deviations and generates outputs that are closer to the references, while still allowing semantically valid paraphrases. This qualitative comparison is consistent with the quantitative improvements observed in the main results. More qualitative analysis is shown in Appendix D.

5 Conclusion

In this paper, we presented CAPO-SLT, a reinforcement learning based framework for stable sign language translation generation. CAPO introduces a confidence-aware clipping strategy for policy optimization. It adapts the upper clipping bound for positive-advantage tokens according to old-policy token probability, while further capping negative-advantage updates to control excessive loss values. This design regulates generation-stage policy updates without changing the underlying SLT backbone. Experiments on Chinese and American sign language benchmarks show that CAPO-SLT consistently improves over corresponding baselines under pose-based settings. Ablations and qualitative examples further support the effectiveness of CAPO for stable and faithful SLT generation.

Limitations

Although CAPO-SLT achieves consistent improvements across Chinese and American SLT benchmarks, the gains on ASL translation are more moderate than those on CSL-Daily. One limitation is that the current reward design is relatively simple and fully reference-based. It combines overlap-oriented metrics with BERTScore, which provides useful sequence-level feedback but may not fully capture valid paraphrases, discourse-level adequacy, or fine-grained semantic preferences in open-ended translation. As a result, the policy optimization signal may be less informative when multiple semantically acceptable translations differ substantially in wording. Future work will explore stronger reward models for SLT generation, such as learned preference models or task-specific semantic evaluators, to provide more reliable feedback for reinforcement learning based SLT optimization.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang,

Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *CVPR*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, pages 10023–10033.

Yuhao Chen, Xin Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, and Xuequan Lu. 2023. Boosting semi-supervised learning by exploiting all unlabeled data. In *CVPR*, pages 7548–7557.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*, pages 5120–5130.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022b. Two-stream network for sign language recognition and translation. *Neurips*, 35:17043–17056.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. 2025. C 2 rl: Content and context representation learning for gloss-free sign language translation and retrieval. *TCSVT*.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized learning assisted with large language model for gloss-free sign language translation. In *LREC-COLING*, pages 7071–7081. ELRA and ICCL.

Helen Cooper, Brian Holt, and Richard Bowden. 2011. Sign language recognition. In *Visual analysis of humans: Looking at People*, pages 539–562. Springer.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *CVPR*, pages 2735–2744.

Edward Fish and Richard Bowden. 2025. Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. *Preprint*, arXiv:2506.00129.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *CVPR*, pages 18362–18372.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jianyuan Guo, Peike Li, and Trevor Cohn. 2025b. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *arXiv preprint arXiv:2505.15438*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *PAMI*, 45(9):11221–11239.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021a. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *CVPR*, pages 11087–11096.

Hezhen Hu, Wengang Zhou, Junfu Pu, and Houqiang Li. 2021b. Global-local enhancement network for nmf-aware sign language recognition. *TOMM*, 17(3):1–19.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with llms. In *NAACL*, pages 3901–3920.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*.

Peiqi Jiao, Yuecong Min, and Xilin Chen. 2024. Visual alignment pre-training for sign language translation. In *ECCV*, pages 349–367. Springer.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Whole-body human pose estimation in the wild. In *ECCV*, pages 196–214. Springer.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. 2025. Leveraging the power of mllms for gloss-free sign language translation. In *ICCV*, pages 21048–21058.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, pages 1459–1469.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. In *ICLR*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *ACL*, pages 12904–12916.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *NAACL*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Neurips*, 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. *ACL*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. In *CVPRW*, pages 5625–5635.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Neurips*, 36:29029–29047.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2025. Signrep: Enhancing self-supervised sign representations. *arXiv preprint arXiv:2503.08529*.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, volume 32.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *CVPR*, pages 2551–2562.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. Best: Bert pre-training for sign language recognition with coupling tokenization. In *AAAI*, volume 37, pages 3597–3605.

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 *ICCV*, pages 20871–20881.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, pages 1316–1325.

Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. 2025. Scaling up multimodal pre-training for sign language understanding. *PAMI*.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *ICCV*, pages 14890–14900.

A More Experimental Studies

A.1 Analysis of Optimization Dynamics

We further compare the optimization dynamics of GRPO and CAPO during RL fine-tuning on CSL-Daily under the same experimental settings. As shown in Fig. 3, both methods keep the updated policy relatively close to the reference policy after an initial adjustment, as reflected by their low KL divergence during most of the training process. This suggests that neither method induces excessively large policy shifts from the supervised initialization.

A more pronounced difference emerges in the entropy dynamics. As shown in Fig. 3(a), the mean entropy under GRPO decreases rapidly at the early stage but gradually rises again during later training, exhibiting a non-monotonic trajectory. This behavior suggests that policy uncertainty increases again later in training, which is consistent with less stable generation-stage optimization. In contrast, Fig. 3(c) shows that CAPO exhibits an early entropy decrease followed by gradual stabilization. The policy exhibits decreasing uncertainty followed by stabilization, without an abrupt entropy collapse or a subsequent increase in uncertainty. Such behavior is particularly desirable for autoregressive SLT generation, where the model should strengthen its preference for high-quality translations while avoiding overly aggressive changes to the output distribution.

Together with the KL curves in Fig. 3(b) and (d), these results provide further evidence that CAPO yields more controlled optimization dynamics than GRPO. While both methods remain close to the reference policy in terms of KL divergence, CAPO maintains a more stable entropy trajectory, consistent with the intended role of confidence-aware clipping in regulating token-level update magnitudes during RL fine-tuning.

A.2 Effect of Cross-modal Alignment

We further examine the role of cross-modal alignment in the supervised initialization stage. As shown in Table 11, both variants use the same CAPO-based RL fine-tuning procedure, and the only difference is whether the cross-modal alignment objective is included before RL. Removing the alignment objective leads to lower performance, with 57.67 B-1, 27.13 B-4, and 56.65 R-L. Adding cross-modal alignment improves the results to 58.82 B-1, 28.68 B-4, and 58.59 R-L. This indicates that a stronger supervised initialization is beneficial for subsequent policy optimization. Importantly, this analysis does not change the conclusion of the main experiments, since all policy optimization comparisons in the main paper use the same supervised initialization. Rather, the result shows that CAPO can benefit from better visual-language representations while remaining a generation-stage policy optimization method.

Table 11: Effect of cross-modal alignment during supervised initialization on the CSL-Daily test set. Both variants use CAPO for RL fine-tuning and differ only in whether the alignment objective is applied.

Cross-modal Alignment	B-1	B-4	R-L
w/o	57.67	27.13	56.65
w/	**58.82**	**28.68**	**58.59**

Table 12: Results on the CSL-Daily test set across three random seeds (42, 43, and 44).

Method	B-1	B-4	R-L
GRPO	$58.48_{\pm 0.00}$	$27.94_{\pm 0.01}$	$57.84_{\pm 0.00}$
DAPO	$58.54_{\pm 0.00}$	$28.14_{\pm 0.01}$	$57.92_{\pm 0.00}$
CAPO	$\mathbf{58.85}_{\pm 0.01}$	$\mathbf{28.65}_{\pm 0.00}$	$\mathbf{58.63}_{\pm 0.01}$

A.3 Results across Different Random Seeds

To further assess the consistency of the improvements brought by CAPO, we repeat the RL fine-tuning experiments on CSL-Daily using three random seeds (42, 43, and 44). All methods are evaluated under the same experimental settings, with only the policy optimization method changed. Table 12 reports the mean and standard deviation over the three runs.

As shown in Table 12, CAPO consistently outperforms GRPO and DAPO across all three metrics while exhibiting very small variance across different random seeds. In particular, compared with DAPO, CAPO improves B-1, B-4, and R-L by 0.31, 0.51, and 0.71 points, respectively. The improvements on B-4 and R-L are especially notable, as these metrics place greater emphasis on longer-range matching and sequence-level consistency. These results indicate that the gains of CAPO are consistent across random seeds rather than arising from a particular training run, further supporting the effectiveness of its confidence-aware policy update.

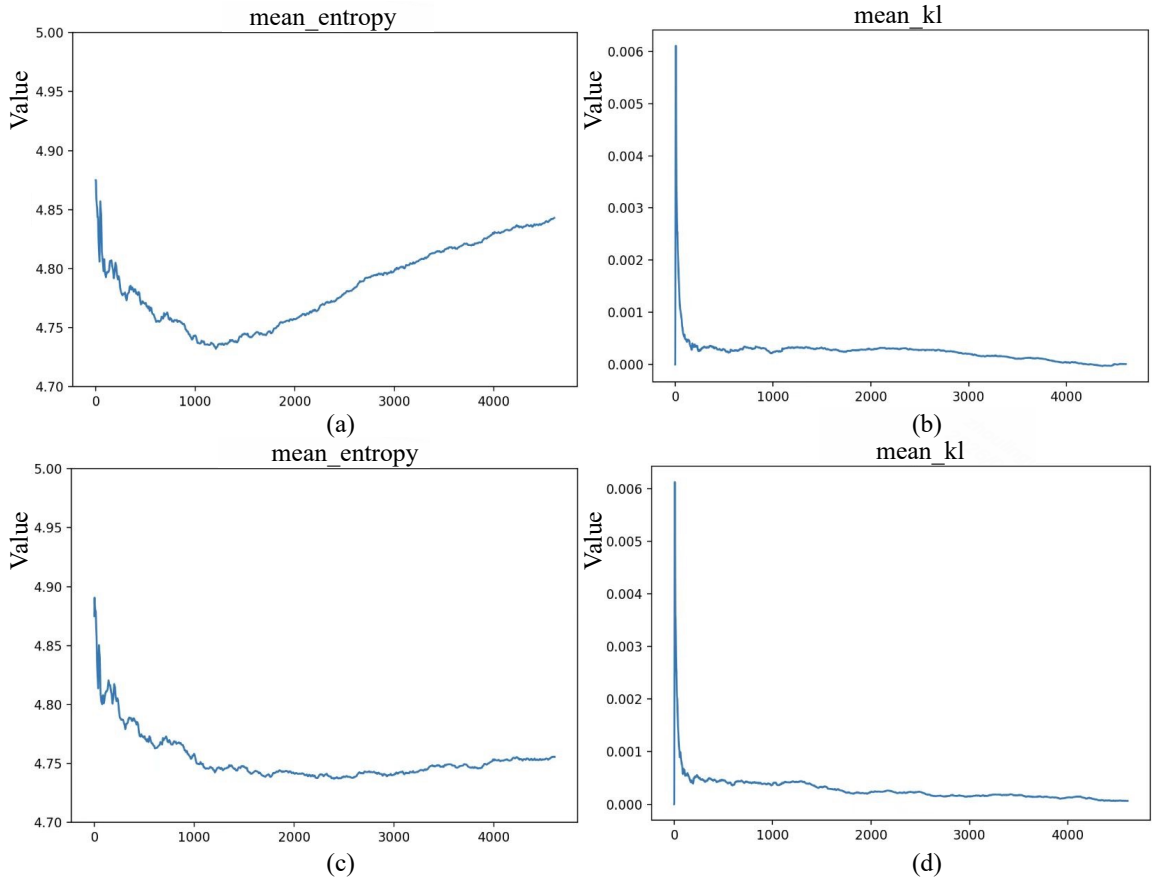


Figure 3: Comparison of training dynamics of GRPO and CAPO during RL fine-tuning on CSL-Daily. (a–b) Mean entropy of the actor policy and mean KL divergence between the current and reference policies during GRPO training, respectively. (c–d) Corresponding results during CAPO training.

B More Experimental Settings

B.1 Details of SLT Datasets

We use five sign language datasets in our experiments, including two downstream SLT benchmarks, two large-scale pre-training corpora, and one isolated sign language recognition benchmark. CSL-Daily (Zhou et al., 2021) is used as the downstream benchmark for Chinese Sign Language translation. It focuses on daily-life communication scenarios such as travel, shopping, medical care, and public services, and provides paired Chinese sentence annotations. Following the official split, CSL-Daily contains 18,401 training segments, 1,077 development segments, and 1,176 test segments, with a total duration of approximately 23 hours. In our experiments, CSL-Daily is used for supervised fine-tuning, reinforcement learning fine-tuning, and final evaluation on Chinese SLT.

How2Sign (Duarte et al., 2021) is used as the downstream benchmark for American Sign Language translation. It is a large-scale multimodal and multiview ASL dataset collected from instructional videos, containing more than 80 hours of sign language videos with aligned English transcripts. Compared with CSL-Daily, How2Sign contains more open-domain instructional content and more diverse English expressions. We use How2Sign to evaluate CAPO-SLT on ASL translation under the same pose-based setting.

For large-scale pre-training, we follow prior SLT studies and use language-specific pre-training corpora before downstream fine-tuning. For Chinese Sign Language, we use CSL-News (Li et al., 2025), a large-scale CSL dataset containing 1,985 hours of sign language videos paired with textual annotations. CSL-News is mainly collected from sign-interpreted news programs and provides a substantially larger source of CSL video-text pairs than downstream benchmarks, making it suitable for learning general visual-language representations before adapting to CSL-Daily. For American Sign Language, we use YouTube-ASL (Uthus et al., 2023), a large-scale open-domain ASL-English corpus mined from YouTube. It contains 984 hours

of captioned ASL videos from 11,093 videos and covers more than 2,500 signers. The broad signer coverage and open-domain content make YouTube-ASL useful for pre-training ASL representations before fine-tuning on How2Sign.

We additionally evaluate isolated sign language recognition on WLASL2000 (Li et al., 2020). WLASL is a large-scale word-level ASL recognition dataset, and its WLASL2000 split covers 2,000 common ASL glosses with over 21,000 video samples from more than 100 signers. Unlike CSL-Daily and How2Sign, WLASL2000 is used for isolated sign recognition rather than sentence-level translation. We include it as an auxiliary benchmark to evaluate whether the learned pose-based representations remain effective for sign understanding beyond continuous translation. The pre-training datasets are only used to initialize the SLT backbone, while downstream results are reported on the official evaluation splits of CSL-Daily, How2Sign, and WLASL2000.

B.2 Training Details

All experiments are implemented in PyTorch (Paszke et al., 2019) and trained on a distributed system comprising eight nodes, each equipped with four NVIDIA L40s GPUs.

Pre-training settings. In this stage, we use the AdamW optimizer with a base learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}, and train the model for 10 epochs with a batch size of 1. The text decoder is instantiated from Qwen3-VL-4B (Bai et al., 2025). We cap the maximum input sequence length at 256 tokens, and fix the feature dimensionality of both visual and textual representations to 512. The instruction template is set as: "You are a translation assistant. Translate sign language video to {Config.lang}." "Config.lang" is set to Chinese or English according to the SLT task. The temperature τ in cross-modal alignment is initialized as $1/0.07$, following common contrastive-learning practice (Radford et al., 2021).

Fine-tuning settings. In this stage, we employ AdamW with a base learning rate of 1×10^{-5} and a weight decay of 1×10^{-6} for all parameters. The batch size is set to 1, and the model is trained for 5 epochs. The model is optimized using the standard cross-entropy loss. For parameter-efficient fine-tuning, we apply LoRA (Hu et al., 2022) to the projection layers of the attention modules, with rank 32 and scaling factor 64. During inference, we use deterministic beam search with a beam size of 5.

Reinforcement learning settings. In this stage, we employ the AdamW optimizer with a base learning rate of 1×10^{-5} and a weight decay of 1×10^{-6} for all parameters. We adopt a new LoRA configuration with rank 4 and scaling factor 4, and train the model for 2 epochs. For CAPO clipping, we set the lower bound δ^- to 0.2, and constrain the adaptive upper clipping range to lie within $[0.2, 0.3]$. During inference, we use deterministic beam search with a beam size of 5.

C Framework Implementation

C.1 Keypoints Extraction

Following prior work (Li et al., 2025), we utilize the RTMPose-x (Jiang et al., 2023) from MMpose to extract keypoints of the whole body. The keypoint definition is based on COCO indices (Jin et al., 2020). All the pose points are divided into four sub-poses: face, body, left hand, and right hand. The indices of each pose are shown as follows:

- **Face**: {24, 26, 28, 30, 32, 34, 36, 38, 40, 54, 84-91},
- **Body**: {1, 4-11},
- **Left hand**: {92-112},
- **Right hand**: {113-133}.

C.2 ST-GCN Architecture

Following prior work (Yan et al., 2018; Fish and Bowden, 2025; Li et al., 2025), each sub-pose group is processed by an ST-GCN stream that can explicitly model the spatial joint relationships and temporal motion dynamics. The core components of the ST-GCN architecture are listed as follows:

1. **Graph Definition:** Each group of keypoints is defined as a graph, and their connections are defined as edges.
2. **Initial Projection:** A linear layer is adopted to project input keypoint coordinates to a higher-dimensional embedding space.
3. **ST-GCN Blocks:** These features are then input into a sequence of ST-GCN blocks, each block consists of a Spatial Graph Convolution (SGC) layer and a Temporal Convolution

Network (TCN). The SGC layer aggregates information from the neighboring joints via a spatial convolution operation, while the TCN layer utilizes 1D convolution across the time dimension to model the motion dynamics.

4. **Residual Connections:** To ensure better feature interaction and stable training, ST-GCN introduces residual connections between spatial fusion layers and temporal fusion layers. This operation allows for richer spatial and temporal relationships.

D More Qualitative Results

We provide additional qualitative results on CSL-Daily to complement the main results. In the following tables, green indicates accurate translations, blue indicates semantically close expressions, and red indicates incorrect, missing, or visually unsupported content. These examples are intended to illustrate representative generation behaviors rather than serve as a substitute for quantitative evaluation.

The additional cases show that CAPO-SLT produces accurate translations across both short and long sentences. For short utterances, the model often matches the reference exactly or differs only in minor function words. For longer sentences, CAPO-SLT maintains coherent event structures and preserves key semantic units such as entities, predicates, temporal expressions, and modifiers. The remaining differences are mostly mild paraphrases or local omissions, rather than large meaning shifts. These examples provide qualitative support for the quantitative results reported in the main paper.

Table 13: Fine-grained error cases on CSL-Daily. The examples show that the remaining errors are mostly local discrepancies, such as missing arguments, temporal mismatches, or minor wording differences, rather than complete sentence-level failures.

Prediction	Reference	Observation
我要 在超市 买椅子， 你去帮我也看看。 I need to buy a chair at the supermarket, can you go and look for me too?	我要去超市买椅子， 你去吗? I need to go to the supermarket to buy a chair, do you want to come?	The main event is preserved, but the motion and question structure are paraphrased into a different intent.
雪下了， 现在 很冷。 It is snowing, and it is very cold now.	下雪了， 今天 很冷。 It is snowing, and it is very cold today.	The translation is semantically close, with a minor temporal expression mismatch.
我预定了一个生日蛋糕。 I ordered a birthday cake.	我 给你 定了一个生日蛋糕。 I ordered a birthday cake for you.	The core action and object are correct, but the beneficiary argument is omitted.
我喜欢冬天 下雪， 很漂亮。 I love snow in winter, it is so beautiful.	我喜欢冬天， 下雪太美了。 I love winter, the snow is so beautiful.	The meaning is largely preserved, while the attachment between season and snow is expressed differently.
弟弟 喜欢踢足球， 身体好。 Younger brother loves playing soccer and is physically strong.	我 弟弟喜欢踢足球，身体 很 好。 My younger brother loves playing soccer and is very physically fit.	The main content is correct, but possessive and degree modifiers are missing.
因为天气不好， 航班取消了。 The flight was canceled due to bad weather.	因为天气不好， 飞机 航班取消了。 The airplane flight was canceled due to bad weather.	The event is correctly translated, with a minor missing nominal modifier.
你公司的午饭好吃吗? Is the lunch at your company good?	你公司的午饭味道怎么样? How does the lunch at your company taste?	The output is a valid paraphrase that preserves the intended question.

Table 14: Additional qualitative examples on CSL-Daily with short sentences. CAPO-SLT generally preserves key entities, attributes, and temporal expressions in short utterances.

Prediction	Reference
今天天气有点冷。 It is a bit cold today.	今天天气有点冷。 It is a bit cold today.
我昨天有点累。 I was a little tired yesterday.	我昨天有点累。 I was a little tired yesterday.
你比我早起一小时。 You woke up an hour earlier than me.	你比我早起一小时。 You woke up an hour earlier than me.
这条裤子怎么样? What do you think of these pants?	这条裤子怎么样? What do you think of these pants?
这位老师教学水平很高。 This teacher has a very high level of teaching ability.	这位老师的教学水平很高。 This teacher has a very high level of teaching ability.
这款游戏有防沉迷系统。 This game has an anti-addiction system.	这款游戏有防沉迷系统。 This game has an anti-addiction system.
我们不想毕业。 We do not want to graduate.	我们不想毕业。 We do not want to graduate.
青春期是身体成熟的时期。 Adolescence is the period when the body matures.	青春期是身体成熟的时期。 Adolescence is the period when the body matures.
要尊重客观规律。 We must respect objective laws.	要尊重客观规律。 We must respect objective laws.
学习是打开知识大门的钥匙。 Learning is the key to unlocking the door to knowledge.	学习是打开知识大门的钥匙。 Learning is the key to unlocking the door to knowledge.
找女朋友是我现在最大的目标。 Finding a girlfriend is my biggest goal right now.	找女朋友是我现在最大的目标。 Finding a girlfriend is my biggest goal right now.
我爱中国共产党。 I love the Communist Party of China.	我爱中国共产党。 I love the Communist Party of China.
会议室里有多媒体系统。 The conference room is equipped with a multimedia system.	会议室里有多媒体系统。 The conference room is equipped with a multimedia system.
你有做生意的头脑。 You have a knack for business.	你有做生意的头脑。 You have a knack for business.