LESS GUESSING! HALLUCINATION-AWARE SELF-CRITICAL TRAINING FOR SIGN LANGUAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the adoption of large-scale pretrained language models (LMs), Sign Language Translation (SLT) has achieved remarkable advances in generating fluent text from sign language videos. However, stronger language generation ability does not necessarily ensure faithfulness to the signed input, and contemporary SLT models may produce fluent translations containing visually unsupported actions or semantically distorted meanings. Despite the serious risks that hallucination poses to the reliability of SLT systems, this problem remains insufficiently studied. In this paper, we systematically investigate hallucination in SLT and characterize it into two forms: Action Hallucination at the visual-content level and Semantic Hallucination at the sentence-meaning level. Based on this characterization, we propose Hallucination-Aware Self-Critical Training (HalluSCT) for Sign Language Translation, a sequence-level optimization framework for hallucination mitigation. HalluSCT introduces a hallucination-aware reward consisting of two components: (i) a Visual Sensitivity Reward that favors translations exhibiting stronger dependence on salient visual evidence under counterfactual perturbation, (ii) a Semantic Consistency Reward that favors translations semantically consistent with the reference translation. The aggregated reward is optimized through self-critical sequence training, encouraging stronger reliance on visual evidence while maintaining semantic correctness. We further introduce SLT-oriented hallucination metrics for evaluating Action Hallucination and Semantic Hallucination. Extensive experiments on public SLT benchmarks, together with human and LLM-based evaluations, demonstrate that HalluSCT consistently reduces hallucination while improving translation quality.

1 INTRODUCTION

Sign Language Translation (SLT) aims to translate continuous sign language videos into natural language sentences. Recent advances, particularly the use of large-scale pretrained language models as translation decoders, have substantially improved fluency and benchmark performance in SLT. However, stronger language generation ability does not necessarily guarantee better faithfulness to the input video. As illustrated in Figure 1, earlier SLT systems often produce disfluent or inaccurate translations, whereas modern LLM-based SLT models may generate fluent and plausible sentences containing information that is partially or entirely unsupported by the sign language. We refer to such outputs as hallucinations in SLT. Based on an empirical analysis of SLT outputs, we identify two representative forms of hallucination, as illustrated in Figure 1. Action Hallucination (AH) occurs when the model generates unsupported or incorrect sign language actions, such as the added "thunderstorms". Semantic Hallucination (SH) occurs when supported sign language content are incorrectly composed, resulting in a sentence-level meaning inconsistent with the reference. SH typical cases include incorrect temporal order or spatial relation, for example, the model assigns sunshine and clouds to the wrong regions. Because these errors are often embedded in fluent sentences, they are difficult to identify and may undermine the reliability of SLT systems in practice.

Hallucination has been extensively studied in large vision-language models (LVLMs). Most existing approaches were initially developed for image-text models, including contrastive alignment Leng et al. (2024); Wang et al. (2024b), visual supervision Liu et al. (2023); Zhang et al. (2024b) and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

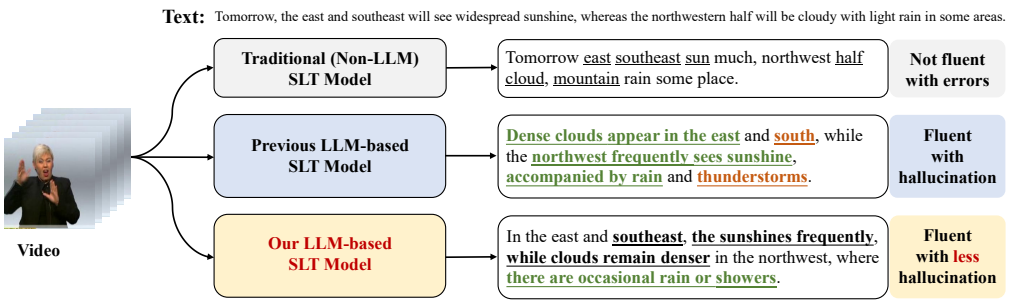


Figure 1: Illustration of hallucination cases from PHOENIX-2014T. The figure presents two types of hallucinations in SLT, where Action Hallucination (AH; the model generates unsupported or incorrect sign language actions) is highlighted in orange, Semantic Hallucination (SH; supported sign language content is incorrectly composed) in green.

reinforcement learning optimization Zhou et al. (2024); Ouali et al. (2024); Xiao et al. (2025); Yang et al. (2025); Zhao et al. (2023); Yu et al. (2024); Yan et al. (2024); Sun et al. (2024). Corresponding evaluation protocols are also largely centered on static visual content, assessing whether generated objects, attributes, or relations are visually supported Rohrbach et al. (2018); Li et al. (2023). Although hallucination research is relatively mature for image captioning, its extension to video understanding remains at an early stage, with recent studies examining dynamic video content such as actions, events, temporal relations, and narratives Li et al. (2025); Choong et al. (2024). Corresponding mitigation strategies include temporal contrastive decoding Zhang et al. (2024a); Wu et al. (2025), visual-feature reweighting Li et al. (2025), and disentangled spatial-temporal representations learning Bae et al. (2025). Nevertheless, these frameworks cannot be directly transferred to SLT. First, in general video understanding, observed action sequences often correspond relatively directly to their textual descriptions. In contrast, SLT involves a non-monotonic mapping from the temporal sequence of sign language expressions to the word sequence of the target spoken language, requiring the model to reorganize the information conveyed in sign language according to the grammar of the target spoken language. When a large-scale pretrained language model is used as the translation decoder, this reformulation process increases the influence of language priors and may introduce content that is fluent but visually unsupported. Second, conventional video models primarily focus on coarse-grained actions, events, and scene dynamics, whereas SLT relies on fine-grained manual and non-manual cues. Small changes in handshape, finger movement, facial expression, or lip movement can alter the lexical meaning or the semantic relations. SLT therefore requires a task-specific characterization of hallucination that captures both fine-grained visual evidence and semantic consistency between the input video and the generated translation.

Despite the reliability risks posed by hallucinated translations, hallucination in SLT remains insufficiently studied. Existing studies mainly examine visual faithfulness, hallucination diagnosis, and the reliability of SLT evaluation metrics Min & Xilin (2023); Hamidullah et al. (2025); Yazdani et al. (2025). However, they do not directly address how to reduce hallucinated generation through sequence-level optimization. Moreover, existing work has not clearly distinguished hallucinations caused by unsupported sign language actions from those caused by incorrect content composition, making it difficult to design targeted mitigation objectives and evaluation metrics. To address these limitations, we propose Hallucination-Aware Self-Critical Training (HalluSCT) for Sign Language Translation, a framework that explicitly mitigates hallucination in SLT. First, based on empirical observations, we characterize two representative forms of hallucination in SLT: Action Hallucination and Semantic Hallucination. Second, we design a hallucination-aware reward consisting of two components: a visual sensitivity reward that encourages translations to rely more strongly on salient visual evidence, and a semantic consistency reward that encourages translations to remain semantically consistent with the reference translation. The resulting reward is then integrated into a self-critical training framework, enabling the SLT policy to reduce hallucinated content without sacrificing translation quality. Third, we propose SLT-oriented hallucination metrics that respectively evaluate Action Hallucination and Semantic Hallucination. We further conduct human and LLM-based evaluations to assess the effectiveness of HalluSCT. We make the following contributions:

- We provide a systematic study of hallucination in SLT and characterize it into two types: Action Hallucination (AH) and Semantic Hallucination (SH).

- We propose Hallucination-Aware Self-Critical Training (HalluSCT) for Sign Language Translation, a sequence-level optimization framework that jointly optimizes a Visual Sensitivity Reward and a Semantic Consistency Reward to reduce hallucination.
- We introduce SLT-oriented hallucination metrics for evaluating the visual evidence support of sign language actions and the semantic consistency of sentences. Extensive experiments on public SLT benchmarks, together with human and LLM-based evaluations, demonstrate that HalluSCT consistently reduces hallucination while improving translation quality.

2 Related Work

Sign Language Translation. Sign Language Translation (SLT) aims to translate sign language videos into natural language sentences. Early studies mainly relied on statistical or recurrent architectures Bungeroth & Ney (2004); Guo et al. (2018); Orbay & Akarun (2020), and were often limited by weak visual representations and language generation capacity. Recent work has increasingly adopted end-to-end neural frameworks with stronger visual encoders, cross-modal interaction modules, and pretrained language decoders Camgoz et al. (2018; 2020b); Li et al. (2020); Zhou et al. (2021); Zhao et al. (2021); Chen et al. (2022b); Kan et al. (2022); Gan et al. (2023); De Coster et al. (2021); Gan et al. (2025), substantially improving translation quality and fluency. However, stronger generation ability does not necessarily ensure faithfulness to the input video. Modern SLT models may still produce fluent translations containing unsupported additions or meaning-distorting errors, making hallucination an important reliability concern.

Recent studies have begun to examine faithfulness and hallucination in SLT, but existing efforts mainly focus on hallucination detection and the limitations of current evaluation protocols Min & Xilin (2023); Hamidullah et al. (2025); Yazdani et al. (2025). How to directly reduce hallucination during sequence-level optimization therefore remains insufficiently explored. In this work, we build on a gloss-based SLT model and investigate hallucination from both mitigation and evaluation perspectives.

Hallucination in LVLMs. Hallucination in large vision-language models (LVLMs) generally refers to generated content that is inconsistent with the visual input. Early studies mainly examined image-based hallucination, particularly unsupported objects, attributes, and relations Rohrbach et al. (2018); Bai et al. (2024). Existing mitigation methods typically strengthen visual supervision or instruction tuning Liu et al. (2023); Zhang et al. (2024b); Sun et al. (2024), contrast model predictions under different visual or decoding conditions Leng et al. (2024); Wang et al. (2024b), or optimize generation using feedback and preference signals **?**Yu et al. (2024); Zhao et al. (2023); Wang et al. (2024a). Corresponding evaluation has evolved from object-centric metrics such as CHAIR to discriminative and multidimensional benchmarks such as POPE and AMBER Rohrbach et al. (2018); Li et al. (2023); Wang et al. (2023).

As LVLMs extend from images to videos, hallucination additionally involves dynamic content and temporal structure. Recent benchmarks evaluate actions, event order and scene transitions through adversarial question answering, caption ranking or paired-video discrimination Wang et al. (2024c); Zhang et al. (2024a); Choong et al. (2024); Li et al. (2025); Rawal et al. (2025). Corresponding mitigation methods employ temporal contrastive decoding Zhang et al. (2024a); Wu et al. (2025), visual-feature reweighting Li et al. (2025), or disentangled spatio-temporal representations Bae et al. (2025). Nevertheless, these approaches are designed primarily for generic video question answering or captioning, where outputs typically describe coarse objects, actions, scenes, or events. SLT instead requires non-monotonic linguistic reformulation and depends on fine-grained manual and non-manual cues. Existing video-oriented approaches therefore do not directly address the two requirements of increasing sensitivity to sign visual evidence and preserving sentence-level semantic consistency, motivating task-specific mitigation and evaluation strategies for SLT.

3 Method

3.1 Overview of HalluSCT

Figure 2 presents the overall framework of HalluSCT. Built upon a supervised gloss-based SLT model, HalluSCT mitigates hallucination by optimizing the translation policy with hallucination-

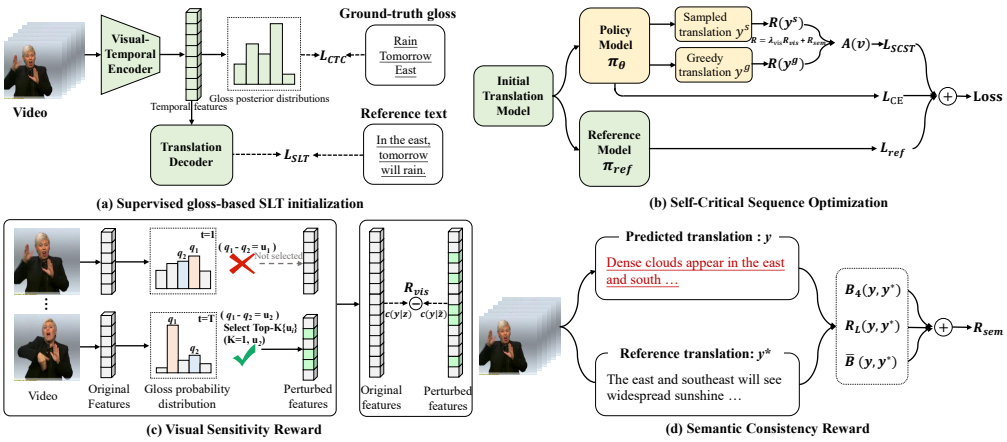


Figure 2: Overview of the proposed HalluSCT framework.

aware rewards. It combines a Visual Sensitivity Reward, which encourages reliance on action-relevant visual evidence to reduce Action Hallucination,with a Semantic Consistency Reward, which encouragess sentence-level agreement with the reference translation to reduce Semantic Hallucination. The combined reward is optimized through self-critical sequence training, with a frozen reference-policy anchor and a supervised translation anchor to stabilize optimization. HalluSCT reduce hallucination while improving translation quality.

3.2 Supervised SLT Initialization

We retain a common gloss-based SLT architecture and mitigate hallucination through a hallucination-aware self-critical training objective rather than architectural modifications. Following TwoStream Chen et al. (2022b); Gan et al. (2025), the architecture combines a gloss recognition branch with a pretrained translation decoder. As illustrated in Figure 2(a), an input sign language video v is processed by a visual encoder and temporal convolutions Li et al. (2020) to produce a temporal feature sequence $\mathbf{z} = \{\mathbf{z}_t\}_{t=1}^{T_v}$, where T_v is the encoded sequence length. A CTC-based recognition branch predicts gloss labels for alignment supervision, and a pretrained translation decoder generates the target language sentence conditioned on the temporal visual features. After supervised training, as illustrated in Figure 2(b), the model is copied into a trainable policy π_θ and a frozen reference policy π_{ref}. The former is updated through hallucination-aware self-critical training, whereas the latter constrains policy updates toward the supervised initialization.

3.3 Hallucination-Aware Reward Design

Our reward design targets the two hallucination types defined above, as illustrated in Figure 2(c) and (d). The Visual Sensitivity Reward encourages candidate translations to depend more on visual evidence, thereby reducing their susceptibility to Action Hallucination. The Semantic Consistency Reward encourages sentence-level agreement with the reference translation to reduce Semantic Hallucination. The two rewards are combined to guide updates of the translation policy.

Visual Sensitivity Reward. Because the temporal order of sign language expressions does not align monotonically with the word order of the target spoken language, reliably assigning each generated token to a specific video segment is difficult. We therefore estimate a translation's dependence on visual evidence through the counterfactual procedure illustrated in Figure 2(c). Specifically, we select high confidence temporal positions according to the gloss posterior distributions, weaken their visual features, and compare the likelihood of the same translation under the original and perturbed feature sequences. A larger likelihood decrease indicates stronger dependence on the selected visual evidence, whereas a smaller decrease suggests that the translation remains likely even after the visual feature is weakened and is therefore more prone to Action Hallucination. This score measures visual dependence rather than directly determining whether a translation is hallucinated, and is therefore used as a training reward.

Given an input video v, the recognition branch produces a temporal feature sequence $\mathbf{z} = \{\mathbf{z}_t\}_{t=1}^{T_v}$ and a CTC posterior distribution over the gloss vocabulary and blank label at each encoded time step. Because the recognition branch is trained with CTC supervision, the margin between its two highest posterior probabilities provides an indicator of recognition confidence. We define the temporal saliency score as

$$u_t = q_t^{(1)} - q_t^{(2)}, \tag{1}$$

where $q_t^{(1)}$ and $q_t^{(2)}$ denote the largest and second-largest CTC posterior probabilities at time step t, respectively. A larger u_t indicates a less ambiguous prediction and is therefore used as a proxy for high confidence sign-recognition evidence. To test whether a generated translation depends on such evidence, we select the k temporal features with the largest confidence margins:

$$\mathcal{I} = \mathrm{TopK}\left(\{u_t\}_{t=1}^{T_v}, k = \lceil \kappa T_v \rceil\right), \tag{2}$$

where $\kappa \in (0, 1]$ is the perturbation ratio and $|\mathcal{I}| = k$. To weaken the selected features while preserving the remaining temporal context, we first compute the mean feature of the encoded video sequence:

$$\bar{\mathbf{z}} = \frac{1}{T_v} \sum_{t=1}^{T_v} \mathbf{z}_t. \tag{3}$$

The perturbed feature sequence $\tilde{\mathbf{z}} = \{\tilde{\mathbf{z}}_t\}_{t=1}^{T_v}$ is defined as

$$\tilde{\mathbf{z}}_t = \begin{cases} \eta \mathbf{z}_t + (1-\eta)\bar{\mathbf{z}}, & t \in \mathcal{I}, \\ \mathbf{z}_t, & t \notin \mathcal{I}, \end{cases} \tag{4}$$

where $\eta \in [0, 1]$ is the feature-retention factor. At each selected position, the perturbed feature retains an η proportion of the original feature and replaces the remaining proportion with the mean feature. For a generated translation $y = (y_1, \ldots, y_N)$, we respectively compute its length-normalized log-likelihood under the original $\mathbf{z}$ and perturbed feature sequences $\tilde{\mathbf{z}}$. This ensures that the likelihood difference is caused by the visual feature perturbation rather than by differences between generated sentences.

$$c(y \mid \mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} \log \pi_\theta \left(y_i \mid y_{<i}, \mathbf{z}\right), \tag{5}$$

$$c(y \mid \tilde{\mathbf{z}}) = \frac{1}{N} \sum_{i=1}^{N} \log \pi_\theta \left(y_i \mid y_{<i}, \tilde{\mathbf{z}}\right). \tag{6}$$

We measure the likelihood reduction caused by weakening the selected features as

$$\Delta_{\mathrm{vis}}\left(y; \mathbf{z}, \tilde{\mathbf{z}}\right) = \max\left(c(y \mid \mathbf{z}) - c(y \mid \tilde{\mathbf{z}}), 0\right). \tag{7}$$

The Visual Sensitivity Reward is then obtained by scaling and clipping this reduction:

$$R_{\mathrm{vis}}\left(y; \mathbf{z}, \tilde{\mathbf{z}}\right) = \mathrm{clip}\left(\frac{\Delta_{\mathrm{vis}}\left(y; \mathbf{z}, \tilde{\mathbf{z}}\right)}{\tau_{\mathrm{vis}}}, 0, 1\right), \tag{8}$$

where τ_{vis} is a scaling constant. A larger reward indicates that weakening the selected high confidence temporal features causes a larger decrease in the likelihood of the translation, suggesting stronger dependence on the selected visual evidence.

Semantic Consistency Reward. Visual sensitivity alone does not ensure that the recognized sign language content is composed into the correct sentence-level meaning. We therefore introduce a reference-based reward as a practical proxy for semantic consistency. For a generated sentence y and its reference translation y^*, we define

$$R_{\mathrm{sem}}(y, y^*) = \omega_4 \, \mathrm{B}_4(y, y^*) + \omega_L \, \mathrm{R}_L(y, y^*) + \omega_{\mathrm{avg}} \, \overline{\mathrm{B}}(y, y^*), \tag{9}$$

where B_4 and R_L denote sentence-level BLEU-4 Papineni et al. (2002) and ROUGE-L Lin (2004), respectively, and $\overline{\mathrm{B}}$ is the average of sentence-level BLEU-1 to BLEU-4. The coefficients ω_4, ω_L, and ω_{avg} balance higher-order phrase precision, sequence-level coverage, and lexical agreement across different n-gram orders.

Overall Reward. The final reward combines visual sensitivity and semantic consistency:

$$R_{\text{total}}(y, \mathbf{z}, \tilde{\mathbf{z}}, y^*) = \lambda_{\text{vis}} R_{\text{vis}}(y; \mathbf{z}, \tilde{\mathbf{z}}) + R_{\text{sem}}(y, y^*), \tag{10}$$

where λ_{vis} controls the contribution of the visual sensitivity reward. The visual reward discourages translations that remain largely unchanged when salient evidence is weakened, Meanwhile, the semantic reward favors translations that remain consistent with the reference, preventing a translation from receiving a high total reward solely because it exhibits strong dependence on visual evidence.

3.4 Self-Critical Sequence Optimization

Self-Critical Training. As illustrated in Figure 2(b), we adopt self-critical sequence training (SCST) Rennie et al. (2017); Gao et al. (2019) to update the policy model according to the hallucination-aware reward. SCST compares a sampled translation with the model's own greedy translation and uses the latter as a baseline, without introducing an additional value model. For each input video v, the policy model π_θ generates a sampled translation $y^s \sim \pi_\theta(\cdot \mid v)$ and a greedy translation y^g. Their reward difference defines the sequence-level advantage:

$$A(v) = R_{\text{total}}(y^s; \mathbf{z}, \tilde{\mathbf{z}}, y^*) - R_{\text{total}}(y^g; \mathbf{z}, \tilde{\mathbf{z}}, y^*). \tag{11}$$

To reduce the influence of sequence length, we use the average token log-probability of the sampled translation:

$$\log \pi_\theta^{\text{avg}}(y^s \mid v) = \frac{1}{|y^s|} \sum_{t=1}^{|y^s|} \log \pi_\theta \left(y_t^s \mid y_{<t}^s, v\right). \tag{12}$$

The self-critical objective is

$$\mathcal{L}_{\text{SCST}} = -A(v)\, \log \pi_\theta^{\text{avg}}(y^s \mid v). \tag{13}$$

This objective increases the likelihood of sampled translations that receive higher hallucination-aware rewards than the greedy baseline, while decreasing the likelihood of lower reward samples.

Stabilized Objective. Optimizing the policy solely with the reward may cause it to deviate from the supervised initialization and degrade translation quality. As illustrated in Figure 2(b), we therefore stabilize policy updates using a frozen reference-policy anchor and a supervised translation anchor. The supervised initialization is retained as a frozen reference policy π_{ref}. For each token in the sampled translation y^s, we penalize the difference between its log-probabilities under the current policy and the reference policy:

$$\mathcal{L}_{\text{ref}} = \frac{1}{|y^s|} \sum_{t=1}^{|y^s|} \left(\log \pi_\theta \left(y_t^s \mid y_{<t}^s, v\right) - \log \pi_{\text{ref}} \left(y_t^s \mid y_{<t}^s, v\right)\right)^2. \tag{14}$$

This reference policy anchor constrains the current policy to remain close to the supervised initialization during self-critical training. As shown in Figure 2(b), we additionally optimize the policy using teacher-forced cross-entropy on the reference translation:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|y^*|} \sum_{t=1}^{|y^*|} \log \pi_\theta \left(y_t^* \mid y_{<t}^*, v\right). \tag{15}$$

This supervised translation anchor preserves the model's ability to generate the ground-truth translation during reward-based optimization. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{SCST}} + \lambda_{\text{ref}} \mathcal{L}_{\text{ref}} + \lambda_{\text{CE}} \mathcal{L}_{\text{CE}}, \tag{16}$$

where λ_{ref} and λ_{CE} control the strengths of the reference-policy and supervised translation anchors, respectively.

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate HalluSCT on two widely used gloss-based SLT benchmarks covering different languages and signing scenarios. PHOENIX14T Camgoz et al. (2018) is a German Sign

Language weather-broadcast dataset containing 7,096 training, 519 development, and 642 test samples, with 1,066 gloss labels and a target vocabulary of 2,877 words. CSL-Daily Zhou et al. (2021) is a Chinese Sign Language dataset collected from daily-life scenarios, containing 18,401 training, 1,077 development, and 1,176 test samples, with 2,000 gloss labels and a target vocabulary of 2,343 words.

Training Details. All self-critical training experiments start from the same supervised initialization. Unless otherwise specified, the Visual Sensitivity Reward uses a perturbation ratio of $\kappa = 0.15$, a feature retention factor of $\eta = 0.20$, and a scaling constant of $\tau_{\text{vis}} = 1.5$. The Semantic Consistency Reward weights are set to $\omega_4 = 0.55$, $\omega_L = 0.35$, and $\omega_{\text{avg}} = 0.10$, and the visual reward coefficient is set to $\lambda_{\text{vis}} = 0.05$. The reference-policy and supervised translation anchor coefficients are set to $\lambda_{\text{ref}} = 0.02$ and $\lambda_{\text{CE}} = 0.10$, respectively. For sampled sequence generation, we use top-p sampling with $p = 0.6$, a temperature of 0.5, and a maximum length of 60 tokens. The policy is optimized using a learning rate of 2×10^{-7}, a batch size of 4, and gradient clipping with a maximum norm of 1.0. We implement the model in PyTorch 2.4.1 and train it on four NVIDIA RTX A6000 GPUs, each with 48 GB of memory, using half-precision training. The visual sensitivity term is linearly warmed up during the first epoch, and training converges within five epochs on both datasets. All other settings follow the supervised pretraining configuration unless otherwise specified.

4.2 Evaluation Metrics

We evaluate model performance from two perspectives: translation quality and hallucination. For standard translation quality, we report ROUGE-L Lin (2004) and BLEU-1/2/3/4 Papineni et al. (2002), which are widely used in SLT to measure lexical and n-gram overlap between predictions and references. We adopt two reference-based metrics to evaluate the two hallucination types. CHAIR evaluates action correctness, and SSC evaluates sentence-level semantic consistency using a frozen multilingual natural language inference (NLI) model.

CHAIR. Following prior SLT hallucination evaluation work Rohrbach et al. (2018); Hamidullah et al. (2025), we adapt instance-level CHAIR by replacing object mentions with normalized content words. Let $n_y(w)$ and $n_{y^*}(w)$ denote the occurrence counts of content word w in the predicted translation y and reference y^*, respectively. We compute

$$\text{CHAIR}(y, y^*) = \frac{\sum_w \max(n_y(w) - n_{y^*}(w), 0)}{\sum_w n_y(w)}, \tag{17}$$

It measures the proportion of predicted content words that are absent from or over-generated relative to the reference. Lower values indicate fewer action hallucination. *Specific language preprocessing is detailed in Appendix.*

SSC. Even when the sign language actions are correctly recognized, the generated sentence may still distort their grammatical meanings or compositional relations. We therefore use a frozen multilingual natural language inference (NLI) model that is independent of policy optimization. Let $p_{\text{ent}}(a, b)$ and $p_{\text{con}}(a, b)$ denote the entailment and contradiction probabilities, respectively, when a is used as the premise and b as the hypothesis. We define $e_{\text{sup}} = p_{\text{ent}}(y^*, y)$ to measure whether the generated translation is supported by the reference, and $e_{\text{cov}} = p_{\text{ent}}(y, y^*)$ to measure whether the generated translation covers the meaning of the reference. Similarly, $c_{\text{sup}} = p_{\text{con}}(y^*, y)$ measures whether the generated translation contradicts the reference, whereas $c_{\text{cov}} = p_{\text{con}}(y, y^*)$ measures whether the reference contradicts the generated translation. We combine bidirectional entailment using the harmonic mean and average the contradiction probabilities:

$$E(y, y^*) = \frac{2 e_{\text{sup}} e_{\text{cov}}}{e_{\text{sup}} + e_{\text{cov}} + \epsilon}, \qquad C(y, y^*) = \frac{c_{\text{sup}} + c_{\text{cov}}}{2}. \tag{18}$$

The final semantic consistency score is

$$\text{SSC}(y, y^*) = \frac{E(y, y^*) + (1 - C(y, y^*))}{2}. \tag{19}$$

A higher SSC indicates stronger bidirectional semantic agreement and fewer semantic hallucinations. *Detailed model and implementation settings are provided in Appendix.*

Table 1: Translation quality comparison on the PHOENIX14T dataset. * indicates methods that use both pose and RGB modalities.

Setting	Method	Dev					Test				
		R	B1	B2	B3	B4	R	B1	B2	B3	B4
Gloss-free	GASLTYin et al. (2023)	–	–	–	–	–	39.86	39.07	26.74	21.86	15.74
	NSLTCamgoz et al. (2018)	31.80	31.87	19.11	13.16	9.94	31.80	32.24	19.03	12.83	9.58
	GFSLT-VLPZhou et al. (2023)	43.72	44.08	33.56	26.74	22.12	42.49	43.71	33.18	26.11	21.44
	Sign2GPTWong et al. (2024)	–	–	–	–	–	48.90	49.54	35.96	28.83	22.52
	SignLLMGong et al. (2024)	47.23	46.88	36.59	29.91	25.25	44.49	45.21	34.78	28.05	23.40
	C^2RLChen et al. (2025)	–	–	–	–	–	50.96	52.81	40.20	32.20	26.75
Gloss-based	Joint-SLRTCamgoz et al. (2020a)	–	45.54	32.60	25.30	20.69	–	45.34	32.31	24.83	20.17
	ConSLTFu et al. (2023)	47.52	–	–	–	24.27	47.65	–	–	–	25.48
	SignBTZhou et al. (2021)	50.29	51.11	37.90	29.80	24.45	49.54	50.80	37.75	29.72	24.32
	SLTUNETZhang et al. (2023)	52.23	–	–	–	27.87	52.11	52.92	41.76	33.99	28.47
	MMTLBChen et al. (2022a)	53.10	53.95	–	–	27.61	52.65	53.97	–	–	28.39
	TwoStream*Chen et al. (2022b)	54.08	54.32	–	–	28.66	53.48	54.90	42.43	34.46	28.95
	HalluSCT	**54.77**	**55.29**	**42.70**	**34.62**	**29.15**	**54.26**	**56.11**	**43.24**	**34.95**	**29.17**

Table 2: Translation quality comparison on the CSL-Daily dataset.* indicates methods that use both pose and RGB modalities.

Setting	Method	Dev					Test				
		R	B1	B2	B3	B4	R	B1	B2	B3	B4
Gloss-free	GASLTYin et al. (2023)	–	–	–	–	–	20.35	19.90	9.94	5.98	4.07
	NSLTCamgoz et al. (2018)	–	–	–	–	–	34.54	34.16	19.57	11.84	7.56
	GFSLT-VLPZhou et al. (2023)	36.70	39.20	25.02	16.35	11.07	36.44	39.37	24.93	16.26	11.00
	Sign2GPTWong et al. (2024)	–	–	–	–	–	42.36	41.75	28.73	20.60	15.40
	SignLLMGong et al. (2024)	39.18	42.45	26.88	17.90	12.23	39.91	39.55	28.13	20.07	15.75
	C^2RLChen et al. (2025)	–	–	–	–	–	48.21	49.32	36.28	27.54	21.61
Gloss-based	Joint-SLRTCamgoz et al. (2020a)	26.75	28.81	16.48	10.78	7.59	26.91	28.70	16.33	10.52	7.29
	ConSLTFu et al. (2023)	41.46	–	–	–	14.80	40.98	–	–	–	14.53
	SignBTZhou et al. (2021)	49.49	51.46	37.23	27.51	20.80	49.31	51.42	37.26	27.76	21.34
	SLTUNETZhang et al. (2023)	53.58	–	–	–	23.99	54.08	54.98	41.44	31.84	25.01
	MMTLBChen et al. (2022a)	53.38	53.81	–	–	24.42	53.25	53.31	–	–	23.92
	TwoStream*Chen et al. (2022b)	55.10	55.21	–	–	25.76	55.72	55.44	42.59	32.87	25.79
	HalluSCT	**54.91**	**58.10**	**44.15**	**33.95**	**26.77**	**55.18**	**58.20**	**44.38**	**34.30**	**27.08**

4.3 Main results

Overall Translation Performance Tables 1 and 2 report the translation results on PHOENIX14T and CSL-Daily. HalluSCT achieves consistently strong performance across both datasets, showing that hallucination-aware optimization does not trade translation quality for faithfulness. On PHOENIX14T, it obtains the best results across all reported metrics on both development and test sets, including a test BLEU-4 of 29.17. On CSL-Daily, HalluSCT achieves the highest BLEU-1 to 4 scores on both splits, improving test BLEU-4 from 25.79 to 27.08 over the strongest gloss-based baseline, while remaining competitive in ROUGE-L. The larger gains in higher-order BLEU scores indicate that HalluSCT improves phrase-level generation and semantic composition rather than merely increasing unigram overlap.

Automatic Hallucination Evaluation. Table 3 evaluates Action Hallucination and Semantic Hallucination using CHAIR and SSC, respectively. HalluSCT achieves the best CHAIR and SSC results across both datasets. Notably, translation quality and hallucination are not necessarily monotonically correlated: on PHOENIX14T, Sign2GPT obtains a higher translation quality than GFSLT-VLP, yet exhibits more severe hallucination. This observation highlights the need for specific hallucination evaluation beyond conventional translation metrics.

4.4 Visual Dependence and Metric Validation

Independent Visual Dependence Evaluation. Adapting the counterfactual reliability analysis and tail-10 pooling of Hamidullah et al. (2025), we use the frozen supervised reference model π_{ref} as a common evaluator. Each translation is rescored under the correct-video, no-video, and mismatched-video conditions. We report mean reliability R_{mean} over all tokens and tail-10 reliability $R_{\text{tail-10}}$ over the least reliable 10% tokens. As shown in Figure 3(a), human-labeled hallucinatory outputs exhibit lower $R_{\text{tail-10}}$ than non-hallucinatory outputs. Figure 3(b) further shows that HalluSCT improves R_{mean} from 0.740 to 0.882 and $R_{\text{tail-10}}$ from 0.232 to 0.460, indicating stronger dependence on the correct video both globally and among the tokens with the weakest visual sup-

Table 3: Hallucination evaluation results on PHOENIX14T and CSL-Daily.

Methods	Dev		Test	
	CHAIR ↓	**SSC ↑**	**CHAIR ↓**	**SSC ↑**
PHOENIX14T				
GFSLT-VLPZhou et al. (2023)	52.42	62.00	53.42	62.20
Sign2GPTWong et al. (2024)	52.84	58.59	53.92	60.38
2S-SLTChen et al. (2022b)	44.39	72.25	44.14	73.36
HalluSCT	**42.66**	**73.59**	**40.91**	**74.92**
CSL-Daily				
GFSLT-VLPZhou et al. (2023)	71.40	46.97	70.61	46.61
Sign2GPTWong et al. (2024)	62.70	52.80	63.46	52.56
2S-SLTChen et al. (2022b)	49.96	74.11	50.42	76.20
HalluSCT	**40.41**	**78.29**	**40.56**	**78.57**

Table 4: Human-based hallucination evaluation on the CSL-Daily test set.

(a) Human and LLM Evaluation

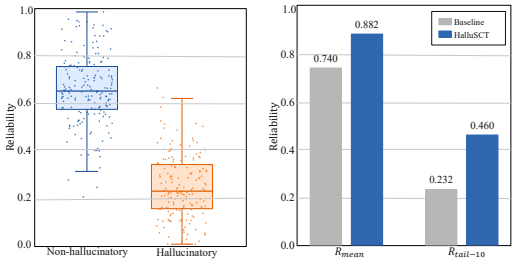
Method	AH ↓	SH ↓	Human-HalRate ↓	LLM-HalRate ↓
Baseline	19.00	53.00	72.00	78.29
HalluSCT	**16.50**	**39.50**	**56.00**	**54.98**

(b) Agreement between automatic scores and human annotations.

Automatic Score	Human Label	AUROC ↑	AUPRC ↑
CHAIR	AH	0.781	0.652
$1 - \mathrm{SSC}$	SH	0.836	0.821
$1 - R_{\text{tail-10}}$	Human-Hal	0.812	0.866

Table 5: Ablation study of reward components on the PHOENIX14T test set.

Method	Reward Components		Evaluation Metrics			
	Vis.	Sem.	R ↑	B4 ↑	CHAIR ↓	SSC ↑
Baseline			52.70	27.15	46.44	70.92
$\text{HalluSCT}_{\text{V}}$	✓		53.67	28.46	43.62	72.83
$\text{HalluSCT}_{\text{S}}$		✓	53.76	28.62	45.40	73.10
$\text{HalluSCT}_{\text{SV}}$	✓	✓	**54.26**	**29.17**	**40.91**	**74.92**



(a) Tail-10 reliability by human hallucination judgment.

(b) R_{mean} and $R_{\text{tail-10}}$ for the baseline and HalluSCT.

Figure 3: Human-grounded visual dependence analysis on CSL-Daily.

port. *Formal definitions, counterfactual construction, and statistical analysis are provided in the Appendix.*

Human and LLM-based Evaluation. To complement the reference-based automatic metrics, we conduct human and LLM-based evaluations on CSL-Daily. We sample 200 CSL-Daily test outputs from the supervised baseline and HalluSCT, respectively, resulting in 400 human-annotated outputs. As shown in Table 4(a), HalluSCT reduces action hallucination from 19.0% to 16.5%, semantic hallucination from 53.0% to 39.5%, and the overall human-judged hallucination rate from 72.0% to 56.0%. The external LLM evaluator reports a consistent reduction from 78.29% to 54.98%. Table 4(b) further shows that the three automatic scores achieve AUROC values above 0.78, with AUPRC values ranging from 0.652 to 0.866, indicating meaningful agreement with human judgments. *Detailed protocols are provided in the Appendix.*

4.5 Ablation Studies

Effect of Reward Components. We evaluate the contributions of the visual sensitivity reward and semantic consistency reward to both translation quality and hallucination reduction, as shown in table 5. Each reward individually improves ROUGE-L, BLEU-4, CHAIR, and SSC over the supervised baseline, showing that enhancing visual dependence and preserving reference semantics are both beneficial. Their combination yields the strongest results on PHOENIX14T, $\text{HalluSCT}_{\text{SV}}$ reduces test CHAIR from 46.44 to 40.91 and increasing SSC to 74.92. It also achieves the highest translation scores, indicating that the two rewards jointly suppress unsupported signed content while improving sentence-level semantic consistency without sacrificing translation quality.

5 Conclusion

In this work, we characterize hallucination in SLT at the signed-action and sentence-meaning levels and propose HalluSCT, a hallucination-aware self-critical training framework. HalluSCT jointly optimizes a Visual Sensitivity Reward and a Semantic Consistency Reward to strengthen reliance on input visual evidence while preserving semantic faithfulness. Experiments using automatic, human, and LLM-based evaluations demonstrate that HalluSCT consistently reduces hallucination while improving translation quality.

REFERENCES

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. Mashvlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13744–13753, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Jan Bungeroth and Hermann Ney. Statistical sign language translation. In *Workshop on representation and processing of sign languages, LREC*, volume 4, pp. 105–108. Citeseer, 2004.

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*, pp. 7784–7793, 2018.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pp. 301–319. Springer, 2020a.

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

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

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

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

Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771*, 2024.

Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. Frozen pretrained transformers for neural sign language translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pp. 88–97, 2021.

Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. A token-level contrastive framework for sign language translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Towards real-time sign language recognition and translation on edge devices. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4502–4512, 2023.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Mixsigngraph: A sign sequence is worth mixed graphs of nodes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. Self-critical n-step training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6300–6308, 2019.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18362–18372, 2024.

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. Hierarchical lstm for sign language translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yasser Hamidullah, Koel Dutta Chowdhury, Yusser Al Ghussin, Shakib Yazdani, Cennet Oguz, Josef van Genabith, and Cristina España-Bonet. Grounding or guessing? visual signals for detecting hallucinations in sign language translation. *arXiv preprint arXiv:2510.18439*, 2025.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. Sign language translation with hierarchical spatio-temporal graph neural network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3367–3376, 2022.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

Yuecong Min and CHEN Xilin. Towards faithful sign language translation. 2023.

Alptekin Orbay and Lale Akarun. Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 222–228. IEEE, 2020.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pp. 395–413. Springer, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. Argus: Hallucination and omission evaluation in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20280–20290, 2025.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*, 2024a.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024b.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024c.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: Leveraging large language models for gloss-free sign language translation. *arXiv preprint arXiv:2405.04164*, 2024.

Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. Season: Mitigating temporal hallucination in video large language models via self-diagnostic contrastive decoding. *arXiv preprint arXiv:2512.04643*, 2025.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25543–25551, 2025.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pp. 37–53. Springer, 2024.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10610–10620, 2025.

Shakib Yazdani, Yasser Hamidullah, Cristina España-Bonet, Eleftherios Avramidis, and Josef van Genabith. A critical study of automatic evaluation in sign language translation. *arXiv preprint arXiv:2510.25434*, 2025.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2551–2562, 2023.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10819–10829, 2022.

Biao Zhang, Mathias Müller, and Rico Sennrich. Sltunet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*, 2023.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024a.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pp. 196–213. Springer, 2024b.

Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, 24:2662–2672, 2021.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

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*, pp. 20871–20881, 2023.

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

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

Appendix of "Less Guessing! Hallucination-Aware Self-Critical Training for Sign Language Translation"

Paper ID 15493

6 AI USE STATEMENT

We used generative AI tools to assist with translation, drafting, and language editing of selected portions of the manuscript. We also sought feedback on the clarity and organization of the methodology section, including the terminology and explanations of model components. The authors reviewed and revised all AI-assisted text and checked the methodological descriptions against the actual implementation. We take full responsibility for the final content of this paper.

7 HALLUCINATION EVALUATION DETAILS

7.1 CHAIR COMPUTATION AND LANGUAGE-SPECIFIC PREPROCESSING

Following Hamidullah et al. (2025), we use the reference translation as a textual proxy for the content conveyed by the sign language video. The prediction and reference are converted into normalized content-token multisets before computing CHAIR. Matching is performed over normalized tokens without embedding-based semantic matching.

German preprocessing. For PHOENIX14T, we use `de_core_news_sm` for tokenization, part-of-speech tagging, and lemmatization. We retain tokens tagged as `NOUN`, `PROPN`, `VERB`, `AUX`, `ADJ`, `ADV`, or `NUM`. Punctuation, spaCy stop words, and the additional German stop words used by the official implementation are removed. The remaining tokens are lowercased, lemmatized, and further normalized with the German Snowball stemmer.

Chinese preprocessing. For CSL-Daily, we first apply Unicode NFKC normalization and convert traditional Chinese characters into simplified Chinese using OpenCC with the `t2s` configuration. Sentences are segmented with jieba, after which punctuation, empty tokens, and the Chinese stop words used by the official implementation are removed. Numerical expressions are retained, and no stemming is applied.

Corpus-level scoring. Let $n_y(w)$ and $n_{y^*}(w)$ denote the occurrence counts of a normalized content token w in prediction y and reference y^*, respectively. The number of unsupported occurrences is

$$h(w) = \left[n_y(w) - n_{y^*}(w)\right]_+. \tag{20}$$

A token is counted as hallucinated when it is absent from the reference or occurs more frequently than supported by the reference. Missing reference tokens are not included in the numerator because CHAIR measures unsupported generation rather than omission. We report the micro corpus-level score

$$\mathrm{CHAIR}_{\mathrm{micro}} = \frac{\sum_i \sum_w h_i(w)}{\sum_i \sum_w n_{y_i}(w)}. \tag{21}$$

If an output contains no retained content token, its sentence-level score is set to zero and it contributes zero tokens to the corpus numerator and denominator. All values reported in the paper are multiplied by 100; lower values indicate fewer Action Hallucinations.

7.2 SSC MODEL AND IMPLEMENTATION

We evaluate sentence-level semantic consistency using the frozen multilingual NLI model `MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`. The model and its tokenizer are used in evaluation mode and are never updated during policy optimization. German and Chinese translations are retained in their original languages. For each prediction–reference pair, the NLI model is

applied in both directions, and a softmax over the NLI logits provides entailment and contradiction probabilities.

Let $p_{\mathrm{ent}}(a, b)$ and $p_{\mathrm{con}}(a, b)$ denote the entailment and contradiction probabilities when a is the premise and b is the hypothesis. We compute

$$e_{\mathrm{sup}} = p_{\mathrm{ent}}(y^*, y), \qquad c_{\mathrm{sup}} = p_{\mathrm{con}}(y^*, y), \tag{22}$$

$$e_{\mathrm{cov}} = p_{\mathrm{ent}}(y, y^*), \qquad c_{\mathrm{cov}} = p_{\mathrm{con}}(y, y^*). \tag{23}$$

The two entailment directions measure whether the prediction is supported by the reference and whether it covers the reference meaning, respectively. Their harmonic mean and the average contradiction probability are

$$E(y, y^*) = \frac{2 e_{\mathrm{sup}} e_{\mathrm{cov}}}{e_{\mathrm{sup}} + e_{\mathrm{cov}} + \epsilon}, \tag{24}$$

$$C(y, y^*) = \frac{c_{\mathrm{sup}} + c_{\mathrm{cov}}}{2}, \tag{25}$$

where ϵ is a small numerical constant. The final score is

$$\mathrm{SSC}(y, y^*) = \frac{E(y, y^*) + 1 - C(y, y^*)}{2}. \tag{26}$$

We average SSC over all evaluation samples and multiply the result by 100. Higher values indicate stronger bidirectional semantic agreement and fewer Semantic Hallucinations. No threshold, model parameter, or calibration parameter is selected using the test set. Because general translation errors or omissions can also lower semantic agreement, SSC is treated as a continuous consistency score rather than a standalone binary hallucination label.

8 Human and LLM-Based Evaluation Protocols

8.1 Human Annotation

We select 200 CSL-Daily test instances and collect translations from the supervised baseline and HalluSCT for the same inputs, yielding 400 system outputs. Each annotation unit consists of a reference translation and one system prediction. The reference serves as the textual proxy for the sign language content, consistent with the reference-based automatic evaluation.

Each output is assessed using two binary labels:

- **Action Hallucination (AH):** the prediction introduces, substitutes, or over-generates a sign language action or content unit that is unsupported by or inconsistent with the reference.
- **Semantic Hallucination (SH):** sign language content is composed into a sentence-level meaning inconsistent with the reference, including incorrect negation, temporal order, or spatial relations.

The two labels are evaluated independently and are not assumed to be mutually exclusive. Pure omissions, stylistic differences, and fluency errors are not marked as hallucinations unless they introduce unsupported content or change the sentence-level meaning. The overall human label, $\mathrm{Human\text{-}Hal}$, is positive when either AH or SH is positive. Human-HalRate is the percentage of outputs with a positive overall label.

8.2 External LLM Evaluation

We additionally use DeepSeek-V3.1 as an external evaluator. For each sample, the evaluator receives the reference translation, the system prediction, and the same AH/SH definitions used for human annotation. It is instructed to judge whether the prediction contains any unsupported action/content or meaning-distorting composition, while not treating omission or surface-level fluency errors alone as hallucination. The evaluator returns a binary overall decision, and $\mathrm{LLM\text{-}HalRate}$ is computed as the proportion of outputs judged to contain hallucination. The LLM evaluator is used only for evaluation and is not involved in HalluSCT training.

8.3 Agreement between Automatic Scores and Human Labels

We evaluate the agreement between continuous automatic risk scores and their corresponding human labels using AUROC and AUPRC. The pairings are

$$\text{CHAIR} \leftrightarrow \text{AH}, \qquad 1 - \text{SSC} \leftrightarrow \text{SH}, \tag{27}$$

$$1 - R_{\text{tail-10}} \leftrightarrow \text{Human-Hal}. \tag{28}$$

AUROC evaluates ranking quality over all possible thresholds, whereas AUPRC emphasizes performance on the positive hallucination class. We compute both metrics directly from the per-output scores and human labels without choosing a threshold on the test set.

9 Independent Visual Dependence Evaluation

9.1 Counterfactual Construction

This evaluation is independent of the HalluSCT reward. We use the frozen supervised reference model π_{ref} as a common evaluator for outputs from both the baseline and HalluSCT. For a fixed candidate translation $y = (y_1, \ldots, y_T)$, we keep the token sequence unchanged and rescore it by teacher forcing under three conditions:

1. the correct input video v;
2. a no-video condition $\varnothing$, in which the visual contribution to the decoder is removed; and
3. a mismatched video v^- that does not correspond to the candidate translation.

The same candidate tokens and prefixes are used in all three passes, so differences in probability arise from the visual condition rather than from different decoded sentences.

For token y_t with prefix $y_{<t}$, define

$$p_{\text{vid},t} = \pi_{\text{ref}}(y_t \mid y_{<t}, v), \tag{29}$$

$$p_{0,t} = \pi_{\text{ref}}(y_t \mid y_{<t}, \varnothing), \tag{30}$$

$$p_{\text{mis},t} = \pi_{\text{ref}}(y_t \mid y_{<t}, v^-). \tag{31}$$

Following the counterfactual principle of Hamidullah et al. (2025), we retain the stronger alternative explanation:

$$p_{\text{cf},t} = \max\{p_{0,t}, p_{\text{mis},t}\}. \tag{32}$$

If either counterfactual explains the token almost as well as the correct video, the token should not be regarded as strongly visually grounded.

9.2 Token and Sentence Reliability

We define the token-level log-probability margin and normalized reliability as

$$m_t = \log p_{\text{vid},t} - \log p_{\text{cf},t}, \tag{33}$$

$$r_t = \sigma(m_t), \tag{34}$$

where $\sigma(\cdot)$ is the sigmoid function. A larger r_t indicates that the correct video provides a stronger probability advantage over both the no-video and mismatched-video conditions. Special tokens and padding positions are excluded from pooling.

We report two sentence-level statistics. Mean reliability averages all valid token scores:

$$R_{\text{mean}} = \frac{1}{T} \sum_{t=1}^{T} r_t. \tag{35}$$

Tail-10 reliability focuses on the least visually supported tokens. Let $\mathcal{B}_{10}$ denote the indices of the $k = \max(1, \lceil 0.1T \rceil)$ smallest token reliabilities. Then

$$R_{\text{tail-10}} = \frac{1}{k} \sum_{t \in \mathcal{B}_{10}} r_t. \tag{36}$$

The corresponding risk score is $1 - R_{\text{tail-10}}$. Mean reliability measures global visual dependence, while tail-10 reliability is sensitive to a small subset of weakly grounded tokens that may be hidden by sequence-level averaging.

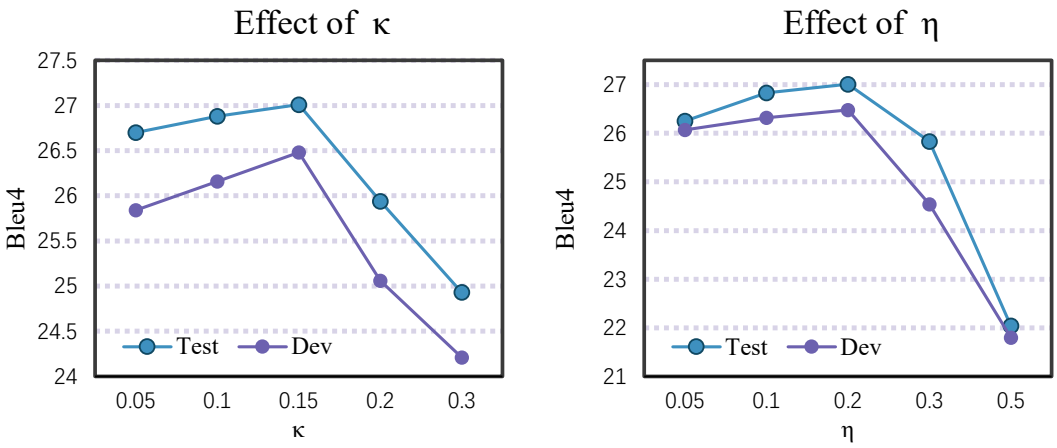


Figure 4: Sensitivity of BLEU-4 to the visual perturbation hyperparameters. Left: perturbation ratio κ. Right: feature-retention factor η.

Table 6: Ablation on visual perturbation strategies used in the visual sensitivity reward on the PHOENIX14T test set.

Perturbation Strategy	R ↑	B4 ↑	CHAIR ↓	SSC ↑
Head Masking	53.35	28.02	48.55	69.71
Hand Masking	53.67	28.46	50.92	70.38
Head and hand Masking	52.34	26.75	48.33	72.54
Salient Feature Perturbation	**54.26**	**29.17**	**40.91**	**74.92**

9.3 Human-Grounded and System-Level Analysis

For the human-grounded analysis, we group the 400 annotated outputs by Human-Hal and compare their per-output $R_{\text{tail-10}}$ values. For the system-level analysis, we macro-average R_{mean} and $R_{\text{tail-10}}$ over test outputs. HalluSCT increases R_{mean} from 0.740 to 0.882 and $R_{\text{tail-10}}$ from 0.232 to 0.460, indicating stronger dependence on the correct visual input both globally and among the least reliable tokens. In addition, the continuous risk $1 - R_{\text{tail-10}}$ is evaluated against Human-Hal using AUROC and AUPRC, as described in Section 8.3. This analysis is descriptive and threshold-free; the frozen evaluator is not fitted or updated on the human labels.

10 Additional Analyses

10.1 Sensitivity to Visual Perturbation Hyperparameters

We study the perturbation ratio κ and feature-retention factor η in the Visual Sensitivity Reward, as shown in Figure 4. The development and test curves exhibit the same overall pattern: BLEU-4 first increases and then decreases as either hyperparameter grows, with the best setting at $\kappa = 0.15$ and $\eta = 0.20$. Insufficient perturbation does not expose dependence on salient visual evidence, whereas overly strong perturbation removes useful signing information and weakens optimization.

10.2 Comparison with Region Masking

We compare salient temporal feature perturbation with three direct region-masking variants on the PHOENIX14T test set, as shown in Table 6. All variants use the same HalluSCT framework and differ only in how the counterfactual visual input is constructed.

Table 7: Controlled temporal perturbation analysis on the PHOENIX14T test set.

Perturbed temporal positions	Confidence gap
Bottom-K	-0.049
Random-K	-0.050
Top-K	0.107

Table 8: Ablation study of reward components on the CSL-Daily test set.

Method	Reward Components		Evaluation Metrics			
	Vis.	Sem.	R ↑	B4 ↑	CHAIR ↓	SSC ↑
Baseline			52.93	24.09	51.55	68.10
HalluSCT$_{\mathrm{V}}$	✓		54.17	25.53	45.15	72.30
HalluSCT$_{\mathrm{S}}$		✓	54.59	26.73	44.23	73.44
HalluSCT$_{\mathrm{SV}}$	✓	✓	**55.18**	**27.08**	**40.56**	**78.57**

Directly masking sign-relevant regions creates an unfavorable trade-off. Hand masking preserves translation quality better than the other masking variants, whereas jointly masking the head and hands improves SSC among the masking baselines but causes the largest BLEU-4 degradation. Salient feature perturbation achieves the best result on all four metrics. It therefore provides a softer intervention that weakens high-confidence temporal evidence without removing entire manual or non-manual regions and without destroying the overall motion context.

10.3 Controlled Saliency Analysis

To verify that the confidence change is associated with salient visual evidence rather than generic distribution shift, we perturb Top-K, random-K, and Bottom-K temporal positions using the same perturbation ratio and strength. The confidence gap is the decrease in the length-normalized likelihood of the fixed translation after perturbation; a larger positive value indicates stronger sensitivity to the perturbed evidence.

As shown in Table 7, top-K perturbation produces a substantially larger confidence gap, whereas random and low-saliency perturbations are nearly identical. This supports the use of the CTC confidence margin as a proxy for locating sign-discriminative temporal evidence and indicates that the reward is not driven solely by arbitrary feature corruption.

10.4 Additional Reward Ablation on CSL-Daily

To complement the reward ablation on PHOENIX14T reported in Table 8, we further evaluate the two reward components on the CSL-Daily test set. The baseline denotes the supervised initialization used for all self-critical training experiments. Both reward components independently improve translation quality over the supervised initialization. The Semantic Consistency Reward yields larger gains in higher-order BLEU scores, while incorporating the Visual Sensitivity Reward further improves the combined model. HalluSCT$_{\mathrm{SV}}$ achieves the best performance across all reported metrics, confirming that the two rewards provide complementary supervision on CSL-Daily.

11 Qualitative Analysis of Hallucination Cases

Table 9 presents qualitative examples from PHOENIX14T and CSL-Daily. We observe that hallucination in SLT can take multiple forms, including incorrect entities, attribute errors, and incorrect semantic relations, even when the generated sentences remain fluent and plausible. Compared with MixSignGraph Gan et al. (2025) and the supervised baseline, HAST-SLT produces outputs that are more faithful to the signed content. These examples further support our quantitative findings that HAST-SLT reduces hallucinations while maintaining translation fluency.

Table 9: Qualitative examples from PHOENIX14T and CSL-Daily. English translations are provided in parentheses below each original sentence. **Action Hallucination** is highlighted in orange, and **Semantic Hallucination** is highlighted in green.

PHOENIX14T dataset	
Case 1	
Ground Truth	morgen im vogtland zwei an rhein und ems bis neun grad (Tomorrow, temperatures will range from two degrees in Vogtland to up to nine degrees along the Rhine and the Ems.)
MixSignGraph Gan et al. (2025)	morgen werden im vogtland zwei grad erwartet, am rhein und an der ems dagegen bis zu zehn grad (Tomorrow, temperatures of two degrees are expected in Vogtland, while along the Rhine and the Ems they will reach up to ten degrees.)
Baseline	morgen temperaturen von zweiundzwanzig grad in der lausitz bis neun grad am rhein (Tomorrow, temperatures will range from twenty-two degrees in Lusatia to nine degrees along the Rhine.)
HalluSCT	morgen temperaturen von zwei grad im vogtland bis neun grad am rhein (Tomorrow, temperatures will range from two degrees in Vogtland to nine degrees along the Rhine.)
Case 2	
Ground Truth	am tag dann regenfälle im süden und südosten am nachmittag örtlich auch gewitter (During the day, there will be rain in the south and southeast, and in the afternoon there will also be isolated thunderstorms.)
MixSignGraph	am tag regnet es im süden und südwesten teils stark am nachmittag einzelne gewitter (During the day, it will rain in the south and southwest, sometimes heavily, with isolated thunderstorms in the afternoon.)
Baseline	am tag regnet es im süden und osten teilweise kräftig am nachmittag einzelne gewitter (During the day, it will rain in the south and the east, partly heavily, with isolated thunderstorms in the afternoon.)
HalluSCT	am tag regnet es gebietsweise im süden und südosten am nachmittag einzelne gewitter (During the day, there will be scattered rain in the south and southeast, with isolated thunderstorms in the afternoon.)
CSL-Daily dataset	
Case 1	
Ground Truth	广告必须首先在内部进行宣传 (The advertisement must first be promoted internally.)
MixSignGraph	广告必须宣传 (The advertisement must be promoted.)
Baseline	必须将这种广告打出去宣传 (This kind of advertisement must be promoted publicly.)
HalluSCT	必须在内部宣传广告 (The advertisement must be promoted internally.)
Case 2	
Ground Truth	对这个故障, 我们有各种应对方法 (For this problem, we have various ways to deal with it.)
MixSignGraph	我们有很多办法避免这次故障发生 (We have many ways to prevent this problem from happening.)
Baseline	我们有不同的办法避免这次故障的发生 (We have different ways to prevent this problem from happening.)
HalluSCT	我们有不同的办法解决这个故障 (We have different ways to solve this problem.)

12 Model Configuration and Limitations

Model configuration. HalluSCT is implemented on a standard gloss-based SLT pipeline. Frame-level visual representations are extracted using PoolFormer-S12 Yu et al. (2022) and processed by temporal convolutions and a bidirectional LSTM. A CTC recognition head provides gloss supervision, while the VL-Embedding module Chen et al. (2022b); Gan et al. (2025) maps encoded visual representations into the embedding space of mBART for autoregressive translation. HalluSCT

changes the sequence-level training objective without adding learnable architectural modules. The supervised model is copied into a trainable policy and a frozen reference policy before self-critical optimization.

Limitations. First, the current experiments focus on gloss-based SLT. Although the self-critical framework is not tied to a particular architecture, extending the Visual Sensitivity Reward to gloss-free models requires an alternative temporal saliency signal, such as pseudo-gloss or model-intrinsic saliency. Second, the Semantic Consistency Reward and the CHAIR/SSC evaluation are reference-dependent and are therefore most suitable for standard supervised benchmarks. Third, visual dependence is operationalized through temporal saliency and counterfactual perturbation; it is a grounding proxy rather than a direct proof that every generated token is aligned with a particular sign segment. Future work should investigate reference-light hallucination evaluation and finer-grained visual grounding for long-range sign language composition.