

Explaining the Influence of Nonverbal Streams Across Sentence Types

Lipisha Chaudhary Ifeoma Nwogu

Department of Computer Science and Engineering

University at Buffalo, Buffalo, NY, USA

{lipishan,inwogu}@buffalo.edu

Abstract

Sign Language Translation (SLT) models integrate multiple visual streams, including the hands, body, and face, to generate written text, yet it remains unclear which streams the model actually relies on. We introduce *Causal Stream Intervention* (CSI), an intervention-based framework that measures each stream’s contribution by suppressing it during decoding and observing the change in the model’s predictions. This goes beyond attention weights, which show where a model *looks* but not what it *depends on*. Across two American Sign Language (ASL) datasets, we find that hand streams drive content-word prediction, while face and body streams contribute in a more variable, context-dependent way, with no single stream consistently dominating during question generation. Attention and routing scores offer a partial signal, but targeted interventions, particularly in later decoding layers, provide a more direct characterization of functional stream use.

1 Introduction

Sign language translation (SLT) maps continuous signing videos into spoken or written language. Unlike spoken language, sign language uses several visual channels at the same time. Hand motion often carries lexical meaning, while facial expression, head movement, and body posture can provide grammatical or sentence-level information (Camgoz et al., 2020a; Gueuwou et al., 2025; Valli et al., 2011). For example, question constructions are often marked not only by the hands, but also by non-manual cues such as facial expression and head movement.

Recent SLT systems increasingly use multiple visual streams, such as hands, body, and face, to improve translation quality (Rust et al., 2024; Li et al., 2025). However, progress is usually measured with sequence-level metrics such as BLEU. These metrics show whether the final translation is close to the reference, but they do not explain which

visual stream supports a specific token prediction. As a result, it remains unclear whether models use linguistically expected cues, such as non-manual markers for questions, or whether correct outputs can be produced mostly from correlations in manual motion.

Attention visualization is often used to interpret multimodal translation models (Ananthanarayanan et al., 2021; Feng et al., 2025), but attention reflects how the model allocates weight, not necessarily what causally affects the output. A stream may receive high attention but have little effect when removed, or it may receive lower attention while still being important for prediction. This issue is especially important in SLT, where meaning is distributed across multiple streams and depends on sentence context.

To address this, we introduce *Causal Stream Intervention* (CSI), an inference-time framework for analyzing stream influence in a multi-stream SLT model. The framework uses separate encoders for the body, face, left hand, and right hand streams, followed by a learned routing mechanism in the decoder. CSI perturbs one stream at a time during decoding and measures the resulting change in token-level negative log-likelihood. This allows us to move beyond asking where the model attends, and instead test which streams actually affect its predictions across sentence types, decoder layers, and targeted intervention settings.

We make the following contributions:

- We introduce a multistream SLT architecture with decoder-side stream gating that supports token-level analysis of stream influence across face, body, left-hand, and right-hand streams.
- We propose an intervention-based evaluation protocol for estimating functional stream influence during decoding, providing a beyond-attention account of streams’ contributions.

- We group results into six coarse sentence categories, with the primary hypothesis-driven analysis focused on questions and declaratives, and benchmark pre-intervention attention/router signals against intervention-measured effects.

Type	Primary Non-Manual Marker
Yes/No Question	Raised brows, forward tilt
Wh-Question	Squinted brows, head tilt
Negation	Side-to-side headshake
Topicalization	Raised brows, head tilt, pause
Conditional	Raised brows, eye gaze shift
Declarative	None

Table 1: Linguistic background on ASL sentence types and typical non-manual markers. These markers motivate the analysis but are not used as direct video annotations; full descriptions and illustrative glosses are in Appendix Table 9.

We test the following hypotheses:

- **H1:** Non-manual (face/body) streams contribute to question token generation, but their functional influence is expected to be context-dependent and distributed across streams.
- **H2:** Manual streams dominate during the generation of declarative and content-bearing tokens.
- **H3:** Attention weights alone are insufficient to characterize stream influence; high attention does not necessarily imply strong causal contribution to token generation.
- **H4:** Stream dependence identified by CSI exhibits layer-wise selectivity: suppressing the intervention-defined dominant stream produces a larger effect than suppressing non-matched streams, particularly in later decoder layers.

2 Background and Related Work

2.1 Continuous multistream SLT modeling

Early continuous SLT work established end-to-end video-to-text learning and benchmarks such as PHOENIX14T (Camgoz et al., 2018), followed by Transformer models that unified recognition and translation (Camgoz et al., 2020b). Signed languages are full natural languages whose computational modeling benefits from linguistically informed treatment of manual and non-manual

structure (Yin et al., 2021). Recent SLT systems have explored large-scale self-supervised pretraining (Rust et al., 2024), gloss-free translation on OpenASL (Lin et al., 2023), and explicit multistream modeling of face, hand, and body inputs (Gueuwou et al., 2025; Damdoo and Kumar, 2025).

Work on sign-language generation further demonstrates that non-manual information is not merely visual detail: prior systems explicitly model prosodic intensification and facial expressions (Inan et al., 2022; Viegas et al., 2023), while more recent generation systems incorporate manual and non-manual ASL components in the synthesis pipeline (Zhang et al., 2025). These works motivate preserving linguistically meaningful streams. CSI differs in objective: rather than improving generation or translation quality, it tests whether a trained stream-aware SLT model functionally relies on the available body, face, and hand evidence during decoding.

2.2 Explainability Beyond Attention

A central debate in NLP interpretability is whether attention weights are valid explanations; prior work argues that attention can be a weak proxy for feature importance and may be altered without changing predictions (Jain and Wallace, 2019). Serrano and Smith (Serrano and Smith, 2019) probe this directly by perturbing attention and find only noisy alignment with true input importance, while Wiegreffe and Pinter (Wiegreffe and Pinter, 2019) emphasize that conclusions depend on the explanation definition and the experimental design. In NMT, Moradi et al. (Moradi et al., 2019) use counterfactual attention interventions to test whether substantially different attention patterns can preserve translations, while Li et al. (Li et al., 2022) probe the contribution of visual features in multimodal machine translation. Transformer intervention methods also manipulate attention or internal activations directly, including AtMan (Deiseroth et al., 2023) and causal tracing/model editing analyses (Meng et al., 2022). Beyond attention, gradient-based attribution methods provide an alternative class of explanation tools (Sundararajan et al., 2017).

2.3 Faithfulness and Attention Critiques

Because qualitative heatmaps are easy to overinterpret, some recent works emphasize faithfulness and robustness of explanations. Adebayo et al. (Adebayo et al., 2018) propose sanity checks (e.g., pa-

parameter randomization and label perturbations) to detect saliency methods that are insensitive to the model or data. Perturbation-based evaluation provides a more causal test of importance by masking evidence and measuring output degradation. These include structured masking via meaningful perturbations (Fong and Vedaldi, 2017) and black-box probing with randomized masks as in RISE (Peters et al., 2018). In NLP, ERASER (DeYoung et al., 2020) standardizes evaluation of rationales using human-annotated evidence spans, thereby helping separate plausibility from faithfulness and encouraging standardized reporting.

3 Causal Stream Intervention (CSI)

In text and vision models, interpretability often studies whether model-internal signals, such as attention or saliency, faithfully explain predictions (Jain and Wallace, 2019; Wiegrefe and Pinter, 2019). In continuous sign language translation, this question is more complex because meaning is distributed across the multiple streams of input (Camgoz et al., 2020a; Gueuwou et al., 2025).

Existing analyses often rely on attention weights or sentence-level metrics. However, attention does not necessarily show causal importance, and BLEU-level changes cannot explain which stream supports a specific target token. To address this, we introduce Causal Stream Intervention (CSI), an inference-time framework that perturbs streams during decoding and measures the resulting change in token-level negative log-likelihood. We use *causal* in the interventional sense of Pearl (2009) and prior work on causal analysis of neural models (Vig et al., 2020). We directly manipulate internal model components and observe the effect on predictions, rather than reading off correlations from attention weights. We do not claim to identify the independent contribution of each articulator, since visual streams are physically correlated in signed input. CSI lets us study not only where the model attends, but which streams, layers, and heads functionally affect its predictions.

3.1 Analysis Setup

Model Architecture. We analyze a Transformer encoder-decoder SLT model designed for controlled stream-level analysis. The goal is not to claim state-of-the-art translation performance, but to study how different visual streams influence decoding. All analysis is performed at inference time

on a trained model with fixed parameters.

The visual input is divided into four streams:

$$\mathcal{M} = \{\text{body, left hand, right hand, face}\}.$$

These streams correspond to different articulatory channels in signed communication. The hand streams mainly represent manual information, while the body and face streams capture non-manual cues that may carry grammatical or sentence-level information. Body and hand inputs are represented as 6D rotations derived from SMPL-X body fits and MANO hand parameters; the face stream uses 3D facial landmarks extracted with RTMPose. All four streams are extracted from the same video at the same frame rate and aligned by frame index, so corresponding stream sequences are frame-synchronous. We deliberately preserve these separately constructed feature streams until decoder-side fusion: this is an architectural choice for controlled analysis, because comparable body/face/hand disentanglement is not readily available once an early-fusion model has collapsed the visual evidence into a shared representation.

For each stream $m \in \mathcal{M}$, the input sequence $X^m = (x_1^m, \dots, x_{L_m}^m)$ is passed through a stream-specific Transformer encoder: $E^m = \text{Enc}^m(X^m)$ where $E^m \in \mathbb{R}^{L_m \times d}$ is the contextualized representation of stream m . All streams use the same encoder architecture, so our analysis focuses on how the decoder uses each stream, rather than on differences between encoder designs.

At decoding step t and decoder layer ℓ , the decoder state $h_t^{(\ell)}$ attends to each stream through a separate cross-attention pathway:

$$z_t^{m,(\ell)} = \text{CrossAttn}^{m,(\ell)}(h_t^{(\ell)}, E^m). \quad (1)$$

Here, $z_t^{m,(\ell)}$ is the information retrieved from stream m for predicting the next token. Since the stream pathways remain separate before fusion, we can analyze each stream’s contribution across layers and token positions. We show the overall architecture in Figure 1.

Token-Level Stream Fusion. After parallel cross-attention, the decoder combines the stream-specific outputs using a learned token-level router, shown in Figure 1 (right). The router takes the concatenation of the four cross-attention outputs, $c_t^{(\ell)} = [z_t^{\text{body},(\ell)}; z_t^{\text{left},(\ell)}; z_t^{\text{right},(\ell)}; z_t^{\text{face},(\ell)}] \in \mathbb{R}^{4d}$, and maps it through a two-layer MLP, $r_t^{(\ell)} =$

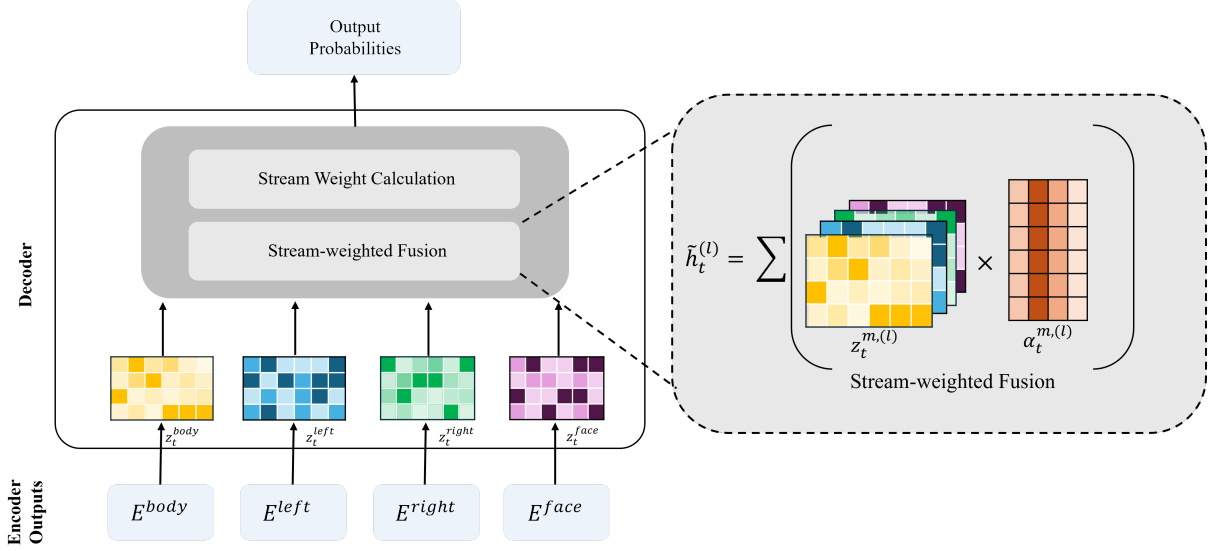


Figure 1: **Stream-aware fusion architecture.** Separate encoders process body, hand, and facial streams, which are combined via learned stream-level weights before decoding.

$W_2 \text{ReLU}(W_1 C_t^{(\ell)} + b_1) + b_2$, where $W_1 \in \mathbb{R}^{d \times 4d}$ and $W_2 \in \mathbb{R}^{4 \times d}$. A softmax over the four logits gives the per-stream weights $\alpha_t^{(\ell)} = \text{softmax}(r_t^{(\ell)})$. For token position t and decoder layer ℓ , the resulting fused representation is:

$$\tilde{h}_t^{(\ell)} = \sum_{m \in \mathcal{M}} \alpha_t^{m,(\ell)} z_t^{m,(\ell)}, \quad \sum_{m \in \mathcal{M}} \alpha_t^{m,(\ell)} = 1. \quad (2)$$

Here, $z_t^{m,(\ell)}$ is the information retrieved from stream m , and $\alpha_t^{m,(\ell)}$ indicates how much weight the decoder gives to that stream when forming the fused representation. The fused cross-attention output is passed through dropout, added to the residual decoder state from the preceding self-attention sublayer, and processed according to the decoder’s configured pre-/post-LayerNorm scheme before the feed-forward sublayer. We use the router weights, together with cross-attention scores, as raw internal signals of stream use. We introduce a routing-based fusion module that dynamically weights each visual stream on a per-token basis, inspired by adaptive routing approaches (Ajirak et al., 2025).

3.2 CSI: Interpreting Stream Contributions

Intervention parameters. Let $p_\theta(y | x, E)$ denote the conditional probability of a target sentence y given the multimodal visual input x and the contextualized stream representations $\{E^m\}_{m \in \mathcal{M}}$, produced by a trained decoder with parameters θ . All token likelihoods are computed by using the reference sequence y as the decoding context. The de-

coder is conditioned on the gold prefix $y_{<t}$ at each step, and the intervention affects only the stream representations attended to during cross-attention. The visual streams enter each decoder layer only through the four stream-specific cross-attention pathways; self-attention and the feed-forward sublayer operate on the decoder state rather than directly on the separate visual representations. This isolates the effect of the intervention on the model’s logit at each reference token, independent of any change in the model’s free-running generation trajectory. For a given intervention $\mathcal{I}$, we compare baseline likelihoods with intervened likelihoods while keeping all model parameters fixed.

For reference token y_t , we compute the baseline and intervened token losses

$$\text{NLL}_{t,\text{base}} = -\log p_\theta(y_t | y_{<t}, x, E), \quad (3)$$

$$\text{NLL}_{t,\mathcal{I}} = -\log p_\theta^\mathcal{I}(y_t | y_{<t}, x, E), \quad (4)$$

and define $\Delta \text{NLL}_t = \text{NLL}_{t,\mathcal{I}} - \text{NLL}_{t,\text{base}}$. Reported table values are means of these token-level effects over the indicated token or sentence subset. A larger positive ΔNLL indicates that suppressing the component reduces probability mass on the reference token. A negative value means that suppression instead increases reference-token confidence, which can occur when a stream supplies noisy, conflicting, or redundant evidence. Because baseline and intervention use the same gold prefix $y_{<t}$, the comparison isolates the intervention from changes in a free-running generated prefix.

Intervention types. We use three types of interventions. First, in *decoder-side stream interventions*, we suppress one stream at a time during cross-attention. At decoder layer ℓ and token position t , we set the selected pre-fusion stream output $z_t^{m,(\ell)}$ to zero while leaving the non-selected stream outputs unchanged. The router is then evaluated normally on the intervened stream representations, allowing its softmax weights to adapt to the suppression.

Second, in *layer-wise targeted interventions*, we apply the same suppression at a specific decoder layer. A token’s dominant stream is computed once from the full decoder-side stream intervention as the stream with the largest ΔNLL , and that label is reused across layer-wise tests. The matched condition suppresses this dominant stream; the nonmatched value averages the three corresponding interventions on the non-dominant streams. Thus, this experiment localizes an already measured stream dependence across decoder layers rather than evaluating a separate pre-intervention predictor.

Third, we use *input-level stream shuffling* as a control. Here, we shuffle the visual features of one stream before encoding. This keeps the input shape valid, but breaks the temporal structure of that stream. Unlike decoder-side interventions, this tests whether the original visual information in the stream is important for decoding.

All interventions are performed at inference time with fixed model parameters.

Stream-level Localization To identify where causal stream dependence appears inside the decoder, we combine internal signals collected before intervention with stream labels obtained after intervention. The goal is not to predict whether the model’s output is correct, but to test whether the decoder contains information about which visual stream is most important for each token.

For each target token, layer, and attention head, we collect baseline features from the frozen model before applying any intervention. We probe these features following prior work on inference-time probing (Ma et al., 2026; Hoscilowicz et al., 2024). These features include attention statistics, router weights, and contribution norms. The target label is the token’s dominant stream, defined as the stream whose perturbation causes the largest increase in token-level NLL.

We train simple linear probes (Li

et al., 2023) to predict the dominant stream label from these baseline features: $\mathbf{p}_\phi(\mathbf{m} \mid \mathbf{x}_{t,\ell,h}) = \text{softmax}(\mathbf{W}\mathbf{x}_{t,\ell,h} + \mathbf{b})$, where $x_{t,\ell,h}$ is the feature vector for token t , layer ℓ , and head h , and $m \in \mathcal{M}$ is one of the four streams. We evaluate probes using attention-only features, router-only features, and their combination.

This localization step tells us where stream-relevant information is most visible inside the decoder (Vig et al., 2020). If a layer or head can predict the dominant stream well, then its baseline features contain information about the stream that the model later depends on under intervention. We use this result to guide the targeted causal interventions in the following experiments.

3.3 Preliminaries

We analyze a multi-stream SLT model with four visual streams: body, left hand, right hand, and face. The left and right hands are treated as *manual* streams, and the body and face as *non-manual* streams. All analysis is token-level.

Top and bottom streams. For each token, we rank the four streams using a chosen score (attention mass, router weight, or ΔNLL). The top and bottom streams are the highest- and lowest-ranked under that score.

Dominant, matched, and non-matched streams. The *dominant* stream for a token is the one whose perturbation causes the largest ΔNLL . An intervention is *matched* if it perturbs the dominant stream, and *non-matched* otherwise.

4 Experiments

We organize the experiments around the four hypotheses. First, we measure stream-level causal effects across sentence types to test whether manual and non-manual streams contribute differently to questions and declaratives. Second, we evaluate whether raw internal signals such as attention and router weights faithfully identify causally important streams. Third, we test targeted stream explanations through layer-wise matched and non-matched interventions. Finally, we use shuffled-label and input-level shuffling controls to verify that the observed effects are not artifacts of the labeling procedure or decoder pathway redundancy.

4.1 Experimental Setup

Datasets. To study stream-level contributions, we focus on American Sign Language (ASL). We

evaluate our framework on two ASL datasets with different scales. OpenASL (Shi et al., 2022) is a large-scale dataset containing approximately 288 hours of video, with 973 samples in the test set. ASLing (Ananthanarayana et al., 2021) is smaller in scale, with 90 linguistically validated test samples. Using both datasets allows us to examine whether our analysis framework remains useful across different dataset sizes and annotation settings. Since our goal is to study stream-level behavior in ASL translation, and due to the availability of linguistically grounded ASL resources, we do not report results on PHOENIX2014T. Baseline translation performance, model configuration, and training details are reported in Appendices E.1 and A.1.

4.2 Causal Stream Effects by Sentence Types

We test whether visual streams contribute differently across sentence types, with the primary analysis focused on questions and declaratives because these categories have the clearest linguistic motivation and the largest usable subsets. We perturb one stream at a time and measure the mean token-level ΔNLL . Positive values indicate reduced reference-token confidence after suppression; negative values indicate that suppression increased reference-token confidence and are retained as signed effects rather than clipped.

For declarative sentences, we report results using both all target tokens and only content-bearing tokens. This is useful because function words and punctuation-like tokens can dilute the effect of manual lexical information when averaged over the full sentence. We therefore define a subset of gloss-like content tokens by removing common function words.

Type of Tokens	B	F	L	R	NM	M	M-NM
ASLing							
Question	-0.008	0.365	0.245	0.313	0.178	0.279	0.101
Decl. tokens	0.173	0.174	0.209	0.225	0.174	0.217	0.043
Content decl.	0.078	0.101	0.401	0.293	0.090	0.347	0.257
OpenASL							
Question	0.134	-0.372	-0.027	0.217	-0.119	0.095	0.214
Decl. tokens	0.176	-0.263	0.018	0.238	-0.043	0.128	0.171
Content decl.	0.133	-0.522	0.014	0.198	-0.194	0.106	0.300

Table 2: **Stream-level causal effects across sentence types.** Values are the mean ΔNLL . B/F/L/R: body, face, left, right hand. M and NM are the manual (L+R) and non-manual (B+F) means.

Visual streams in signed input are not indepen-

dent. The left and right hands are coupled during two-handed signs, and body posture and facial expression co-vary during non-manual marking. As a result, perturbing a single stream does not cleanly isolate that articulator’s contribution, since correlated information may remain in the other streams. To address this within-group correlation, we report aggregate *Manual* (M) and *Non-Manual* (NM) effects as our primary comparison, computed as the mean ΔNLL over the left/right-hand streams and over the body/face streams, respectively. Individual stream effects are reported for completeness.

Table 2 reports stream-level effects for questions, all declarative tokens, and content-bearing declarative tokens.

4.3 Faithfulness of Raw Internal Signals

After measuring stream-level causal effects, we next test whether the model’s raw internal signals can identify the streams it depends on during decoding. We focus on two signals available before any intervention: cross-attention scores and stream-level router weights.

Signal	Top ΔNLL	Bottom ΔNLL	$\Delta\text{BLEU} \downarrow$	$\Delta\text{ROUGE-L} \downarrow$	$\Delta\text{METEOR} \downarrow$
ASLing					
Attention	0.360	0.139	7.15 / 2.05	11.7 / 2.4	9.2 / 1.8
Router	0.399	0.164	8.85 / 2.22	11.6 / 4.1	8.8 / 3.0
OpenASL					
Attention	0.258	-0.211	2.76 / 0.05	6.4 / 0.9	5.8 / 0.4
Router	0.190	0.064	1.59 / 2.10	5.1 / 4.5	4.9 / 3.6

Table 3: **Faithfulness of raw stream-ranking signals.** Top/Bottom ΔNLL are measured after ablating the highest- and lowest-ranked streams. For each sentence metric $q \in \{\text{BLEU}, \text{ROUGE-L}, \text{METEOR}\}$, we compute $\Delta q = q_{\text{base}} - q_{\text{int}}$; table entries are reported as $\Delta q_{\text{top}} / \Delta q_{\text{bottom}}$.

For each example, we use these signals to rank the four streams: body, face, left hand, and right hand. We then ablate the highest-ranked stream and the lowest-ranked stream under each signal and measure the change in token-level NLL. If a signal is faithful, removing its top-ranked stream should increase NLL more than removing its bottom-ranked stream. In this experiment, ΔNLL is used only as the causal outcome of the intervention, not as a signal for ranking the streams.

Table 3 reports the mean ΔNLL after ablating the top- and bottom-ranked streams for each raw signal. We also report sentence-level performance changes under the same top- and bottom-ranked

ablations to compare token-level causal effects with standard generation metrics.

Dataset	Layer	Nonmatch Δ NLL	Match Δ NLL	Diff Δ NLL
ASLing	0	0.009	0.046	0.037
	1	-0.005	0.059	0.064
	2	-0.003	0.108	0.111
	3	-0.024	0.116	0.140
	4	-0.043	0.292	0.335
	5	-0.059	0.279	0.338
OpenASL	0	-0.021	0.022	0.043
	1	-0.030	0.068	0.097
	2	-0.045	0.090	0.134
	3	-0.038	0.144	0.182
	4	-0.078	0.186	0.264
	5	-0.127	0.214	0.341

Table 4: **Layer-wise causal selectivity.** Matched interventions perturb the token’s dominant stream; non-matched perturb other streams. Diff is matched minus nonmatched Δ NLL.

4.4 Layer-wise Targeted Interventions

We next use token-specific stream labels to test where stream information is most relevant inside the decoder. For each target token, we define its dominant stream as the stream whose perturbation causes the largest increase in token-level NLL.

Using these dominant stream labels, we perform targeted interventions at each decoder layer. A matched intervention perturbs the token’s dominant stream, while a nonmatched intervention perturbs one of the other streams. For example, if a token is assigned to the face stream, then perturbing the face representation is treated as matched, while perturbing the body, left-hand, or right-hand representation is treated as nonmatched.

Table 4 reports the mean Δ NLL for matched and nonmatched interventions across decoder layers. This setup allows us to compare whether perturbing the labeled stream has a different effect from perturbing streams not assigned to the token.

Dataset	Layer 0	Layer 5
ASLing	0.037 [0.026, 0.049]	0.338 [0.297, 0.381]
OpenASL	0.043 [0.041, 0.046]	0.341 [0.331, 0.351]

Table 5: Bootstrap 95% confidence intervals for the matched–nonmatched Δ NLL gap at the first and final decoder layers. Values are reported as gap [95% CI].

Bootstrap 95% confidence intervals for the matched–nonmatched gap are strictly positive at

every decoder layer in both datasets, supporting the robustness of the layer-wise selectivity pattern. Representative intervals for the first and final decoder layers are reported in Table 5.

Layer	ASLing		OpenASL	
	Original	Shuffled	Original	Shuffled
0	0.037	-0.003	0.043	0.003
1	0.064	0.007	0.097	0.009
2	0.111	-0.004	0.134	0.009
3	0.140	-0.009	0.182	0.011
4	0.335	-0.019	0.264	0.016
5	0.338	0.042	0.341	0.024

Table 6: **Shuffled-label control.** Original is the matched-nonmatched gap with true labels; Shuffled is the same gap after permuting token labels within each layer.

4.5 Shuffle-Based Intervention

Shuffle-based interventions serve as controls, testing whether stream effects are artifacts of labeling or the decoder pathway.

Shuffled-Label Control To verify that the targeted stream labels described in Section 3.2 are not arbitrary, we compare them with a shuffled-label control. In this setting, we randomly shuffle the dominant-stream labels within each layer and repeat the matched-versus-nonmatched intervention analysis. This intervention checks whether the results change when tokens are no longer paired with their original stream labels. If the labels are arbitrary, shuffling them should produce a similar matched/nonmatched gap. If the labels capture meaningful stream assignments, shuffling should reduce this gap. Table 6 reports the matched/nonmatched gap for the original labels and the shuffled labels across decoder layers.

Dataset	Type	Metric	Body	Left	Right	Face	All
ASLing	Ques.	BLEU	20.88	11.09	6.97	16.86	0.84
		Δ NLL	0.495	0.291	0.827	0.045	1.648
	Decl.	BLEU	7.98	10.96	7.93	13.07	4.58
		Δ NLL	0.950	0.337	0.705	0.073	2.186
OpenASL	Ques.	BLEU	13.67	14.65	12.48	16.41	6.25
		Δ NLL	0.196	0.061	0.267	0.000	0.734
	Decl.	BLEU	15.08	17.78	13.29	19.40	5.46
		Δ NLL	0.234	0.124	0.300	0.000	0.866

Table 7: **Input-level shuffle ablation by sentence type.** BLEU is computed after shuffling each stream; Δ NLL is the mean change during decoding. Blue cells mark the most-affected manual (L/R) stream; orange cells mark the most-affected non-manual (B/F) stream.

Input-Level Stream Shuffling Finally, we add an input-level stream shuffling experiment to complement the decoder-side interventions. In the previous experiments, we perturbed stream representations inside the decoder. Here, we instead perturb the visual input before it enters the model.

For each stream, we randomly shuffle its visual features while keeping the tensor shape unchanged. This keeps the input size valid for the model, but disrupts the temporal order and frame-to-frame consistency of that stream. We apply this procedure to each stream separately, and also to all streams together. This experiment helps us test whether the model relies on the structured information within each visual stream, rather than only on the presence of the stream itself. Table 7 reports the effect of input-level stream shuffling for each sentence type and stream.

5 Analysis and Discussion

5.1 Stream Influence Is Distributed

The stream-level results show that the model uses both manual and non-manual cues, rather than relying on a single stream for each sentence type. For declarative sentences, manual streams have stronger effects, especially for content-bearing tokens. This pattern holds in both datasets, where the M–NM gap grows from 0.043 to 0.257 on ASLing and from 0.171 to 0.300 on OpenASL when we restrict to content tokens. This supports H2 and matches the expectation that hands carry much of the lexical information needed for translation.

For question-type sentences, the pattern is more mixed, and the evidence for H1 is weaker. Although questions are often associated with facial and body cues, the model does not rely only on non-manual streams. ASLing shows a strong face effect for questions, while OpenASL shows stronger effects from the right hand and body. The input-level shuffling results also show that disrupting manual and body streams can affect question decoding, while shuffling the face stream alone does not always cause the largest change.

Non-manual question marking is typically spread across many frames and overlaps with manual signing, so its effect may not show up as a strong signal on a single token. Overall, stream influence in SLT is distributed across manual and non-manual cues. Negative ΔNLL values are informative signed effects rather than failures of the intervention: they indicate cases where removing a

stream increases reference-token confidence, consistent with noisy, conflicting, or redundant evidence in correlated stream representations.

5.2 Limits of Raw Internal Signals

The faithfulness results show that internal signal scores are useful, but not complete, indicators of stream importance. Prior work argues that explanations should reflect the model’s actual decision process, not only appear plausible (Jacovi and Goldberg, 2020). In our results, ablating the top-ranked stream usually causes a larger ΔNLL than ablating the bottom-ranked stream, suggesting that these signals capture some information about which streams the model uses (Table 3).

However, this pattern is not consistent across datasets, metrics, or sentence types. A stream may receive high attention or a high router score without producing the strongest causal effect when perturbed. For example, attention gives a larger top–bottom ΔNLL gap than router scores on OpenASL, while router scores are stronger on ASLing. The sentence-type breakdown in Table 12 in the Appendix shows similar variation. This agrees with prior work showing that explanation scores need careful validation before being treated as faithful (Hsia et al., 2024; Paul et al., 2024).

Overall, raw internal scores can indicate where the model assigns attention or routing weight, but they do not fully show which stream the model depends on. This motivates intervention-based analysis, where stream importance is tested by measuring prediction changes after perturbation.

5.3 Targeted Interventions Localize Stream Dependence

The layer-wise targeted intervention analysis asks where an intervention-defined stream dependence becomes most selective inside the decoder. Dominant-stream labels are obtained from the full stream intervention and then reused for layer-specific suppression; therefore, this analysis is an oracle-conditioned localization test rather than an independent pre-intervention faithfulness test. The independent faithfulness comparison is Table 3, which ranks streams using attention or router scores available before intervention.

From Table 4, this effect becomes stronger in later decoder layers. In both datasets, the matched–nonmatched gap is small in early layers and largest near the final layers. This suggests that stream-specific influence becomes clearer as the decoder

gets closer to producing the output token.

The shuffled-label control supports the localization interpretation. When the dominant-stream labels are randomly shuffled, the matched–nonmatched gap becomes much smaller, showing that the layer-wise selectivity depends on preserving the token–stream pairing rather than merely suppressing any stream.

Together with the pre-intervention comparison in Table 3, these results support H4: attention and router scores provide only partial evidence of functional dependence, while CSI directly measures stream-specific prediction changes and shows where those effects become strongest across layers.

5.4 Linguistic Interpretation at Token-Level

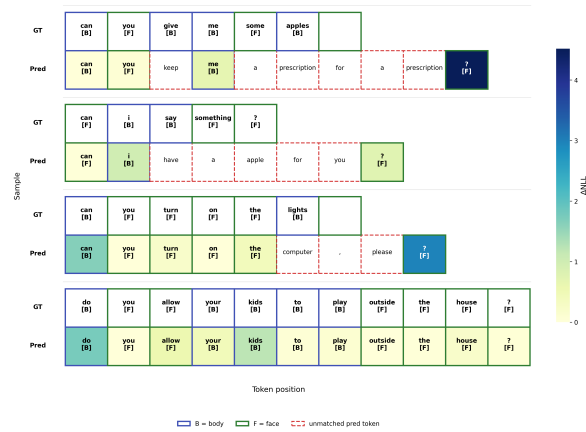


Figure 2: Most affected tokens in a Question-type sentence. Interrogative symbols like “?” cause the most disruption in the model behavior. *Examples from ASLing. An enlarged version is provided as Figure 3 in the Appendix for better readability*

The token-level results show that stream influence in SLT should be interpreted in context. A sentence type may be associated with a particular cue, such as a facial expression for questions, but the model can still rely on several streams when generating the translation. This is expected because the sign language meaning is built from manual and non-manual cues working together, discussed in Appendix C.

Figure 2 gives a qualitative example. In the question-type sentence, question-related markers such as “?” are among the most affected tokens. This suggests that non-manual information can influence how the model handles interrogative structure. At the same time, the sentence is still decoded using both manual and non-manual evidence, rather

than facial cues alone.

Overall, this analysis reinforces the main finding that stream influence is distributed and context-dependent. Intervention-based analysis reveals this behavior by measuring how each stream affects predictions, not just where the model attends.

6 Conclusion

We presented Causal Stream Intervention (CSI), an inference-time framework for measuring how individual visual streams influence token prediction in multi-stream sign language translation. By directly perturbing stream representations during decoding, CSI moves beyond attention and routing scores, which we show provide only partial evidence of functional stream use. Layer-wise targeted interventions show that intervention-defined stream dependence becomes more selective in later decoder layers.

Across two ASL datasets, the model relies on a distributed mix of manual and non-manual cues rather than a single dominant stream per sentence type. Manual streams clearly drive content-bearing tokens in declaratives, but for questions, the linguistically expected dominance of non-manual cues does not hold cleanly. This gap between linguistic theory and learned model behavior is itself informative: it suggests that current multi-stream SLT models may not be using the non-manual evidence that ASL grammar most depends on, even when that information is available in the input. CSI offers a way to surface these mismatches systematically, and we hope it can serve as a diagnostic tool for future multi-stream SLT systems.

7 Limitations

Our analysis is limited to ASL datasets and a controlled multi-stream SLT architecture with explicit body, hand, and face pathways. This design makes stream-level intervention possible, but the findings may not directly transfer to models where visual information is fused into a single representation. Sentence-type labels are derived from reference translations rather than full linguistic annotation of the signed videos, so the sentence-type analysis is coarse-grained.

Visual streams in signed input are physically correlated: the hands often move together during two-handed signs, and body posture co-varies with facial expression. Our interventions therefore do not isolate the independent contribution of each articulator, which is why we report aggregate Manual and Non-Manual effects as the primary comparison. Individual stream effects should be interpreted with this within-group correlation in mind.

Δ NLL is computed under teacher forcing against the reference sequence, isolating the effect of stream perturbation on reference-token probability mass. This is a clean measurement target but differs from free-running generation, where the model’s prediction trajectory may diverge from the reference. The stream-importance patterns we report should therefore be read as evidence about how the model uses its streams when conditioned on the gold prefix.

References

- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *NIPS*, NIPS’18, page 9525–9536. Curran Associates Inc.
- Marzieh Ajirak, Oded Bein, Ellen Rose Bowen, Dora Kanellopoulos, Avital Falk, Faith M Gunning, Nili Solomonov, and Logan Grosenick. 2025. Learning to route: Per-sample adaptive routing for multimodal multitask prediction. *NeurIPS*.
- Tejaswini Ananthanarayana, Nikunj Kotecha, Priyanshu Srivastava, Lipisha Chaudhary, Nicholas Wilkins, and Ifeoma Nwogu. 2021. Dynamic cross-feature fusion for american sign language translation. In *FG 2021*, pages 1–8. IEEE.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. MIT Press.
- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.
- Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. In *Proceedings of ECCV Workshops*.
- Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, pages 10023–10033, Virtual. IEEE/CVF.
- Rina Damdoo and Praveen Kumar. 2025. [Enhanced sign language translation using three stream multimodal occlusion resilient transformer](#). *Discover Computing*, 28:260.
- Bjorn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2023. Atman: Understanding transformer predictions through memory efficient attention manipulation. In *Advances in Neural Information Processing Systems*, volume 36.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL*, pages 4443–4458, Online. Association for Computational Linguistics.
- Yi Feng, Chuanyi Li, Jiatong He, Zhenyu Hou, and Vincent Ng. 2025. Multimodal neural machine translation: A survey of the state of the art. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.
- Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437.
- Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025. Signmusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521.

- Jakub Hoscilowicz, Adam Wiacek, Jan Chojnacki, Adam Cieslak, Leszek Michon, Vitalii Urbanevych, and Artur Janicki. 2024. [Non-linear inference time intervention: Improving llm truthfulness](#). *Preprint*, arXiv:2403.18680.
- Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary Lipton. 2024. [Goodhart’s law applies to NLP’s explanation benchmarks](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1322–1335, St. Julian’s, Malta. Association for Computational Linguistics.
- Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. [Modeling intensification for sign language generation: A computational approach](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *CoRR*, abs/1902.10186.
- Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. In *NIPS*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.
- Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *J. Mach. Learn. Res.*, 25(1).
- Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. [On vision features in multimodal machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Ze Cheng 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, Barcelona, Spain. Association for Computational Linguistics.
- Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. [Gloss-free end-to-end sign language translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.
- Xueqi Ma, Shuo Yang, Yanbei Jiang, Shu Liu, Zhenzhen Liu, Jiayang Ao, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. 2026. [Attention in space: Functional roles of vlm heads for spatial reasoning](#). *Preprint*, arXiv:2603.20662.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35.
- Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2019. Interrogating the explanatory power of attention in neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Hong Kong. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. [Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.
- Judea Pearl. 2009. *Causality*, 2 edition. Cambridge University Press.
- Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 151. BMVA Press.
- Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.
- Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

- Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *EMNLP*.
- William C. Stokoe. 1960. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. University of Buffalo.
- Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *EMNLP*, Brussels, Belgium. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 3319–3328. JMLR.org.
- Clayton Valli, Ceil Lucas, Kristin J. Mulrooney, and Miako Villanueva. 2011. *Linguistics of American Sign Language: An Introduction*, 5 edition. Gallaudet University Press.
- Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. 2023. [Including facial expressions in contextual embeddings for sign language generation](#). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.
- Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *EMNLP-IJCNLP*, Hong Kong, China. Association for Computational Linguistics.
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.
- Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. 2025. [Towards ai-driven sign language generation with non-manual markers](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

A Model, Training & Implementation Details

A.1 Model architecture.

We use a Transformer encoder–decoder model designed for controlled stream-level analysis. The input is divided into four visual streams: body, left hand, right hand, and face. Each stream is processed by a separate Transformer encoder with the same architecture but independent parameters. This gives us one encoded representation per stream, which allows us to analyze how the decoder uses each visual source during translation. Body and hand features are obtained from SMPL-X body fits with MANO hand parameters and converted to 6D rotation representations; the face stream uses 3D facial landmarks extracted with RTMPose. All streams are extracted from the same video at the same frame rate and aligned by frame index.

At each decoder layer, the decoder attends to the four encoded streams through separate cross-attention pathways. The router concatenates the four d -dimensional cross-attention outputs into a $4d$ -dimensional vector, projects it to d dimensions with a linear layer followed by ReLU, and then projects to four router logits. A softmax over those logits yields one weight for each stream, and the weighted stream outputs are summed; dropout is applied to the fused representation. Because the stream pathways remain separate before this fusion, CSI sets the selected pre-fusion stream output to zero while leaving the other stream outputs unchanged; the router then recomputes its weights normally from the intervened stream representations. This separation is what makes the model suitable for controlled stream-level analysis.

The model is trained normally for SLT using teacher-forced sequence-to-sequence cross-entropy loss. After training, all parameters are frozen, and all CSI experiments are performed only at inference time. We do not update the model during intervention analysis.

A.2 Training & Implementation Details

All experiments are implemented in PyTorch using the Fairseq framework. We use a Transformer encoder–decoder model with separate visual encoders for the body, left hand, right hand, and face streams. These streams are fused in the decoder through cross-attention and a token-level routing mechanism.

The encoder and decoder embedding dimensions

are set to 1024 and 300, respectively. The model uses 4 encoder attention heads and 10 decoder attention heads. We train with the Adam optimizer using a learning rate of 1×10^{-4} , label-smoothed cross-entropy loss with smoothing 0.1, gradient clipping with norm 1.0, and dropout across layers. A reduce-on-plateau scheduler is used, with early stopping based on validation performance.

To stabilize training, routing and stream masking are introduced after an initial warm-up period. Specifically, router parameters are activated after 50 updates, allowing the base model to first learn aligned multimodal representations. Models are trained for up to 200 epochs, with validation after each epoch. During evaluation, sequence generation uses beam search with beam size 10.

All CSI analyses, including cross-attention inspection, router-weight analysis, and stream ablations, are performed only at inference time using the best-performing checkpoint. No additional fine-tuning or retraining is applied during the analysis.

A.3 Evaluation Metrics

We report standard NLP translation metrics to measure how interventions affect the final generated sentence. Specifically, we use BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). BLEU measures n-gram overlap between the generated sentence and the reference, ROUGE-L captures the longest common subsequence, and METEOR uses more flexible word-level matching. These metrics provide a sentence-level view of translation quality and make our results comparable to prior SLT and machine translation work.

However, these metrics alone are not enough for stream-level analysis. They can show that the final translation changes after an intervention, but they cannot identify which visual stream supported a specific token prediction. For this reason, we also use token-level negative log-likelihood (NLL). By comparing NLL before and after perturbing a stream, we compute ΔNLL and measure how much that stream affects the model’s confidence in the correct token. Thus, BLEU, ROUGE-L, and METEOR evaluate surface-level translation quality, while ΔNLL provides a finer-grained measure of causal stream influence during decoding.

Component	Configuration
Input streams	Body, left hand, right hand, face
Input features	6D rotations for body/hands; 3D facial landmarks
Encoder layers	[1] per stream
Decoder layers	[6]
Encoder embedding dim	[1024]
Decoder embedding dim	[300]
Decoder Attention heads	[10]
Fusion module	Token-level stream router
Training objective	Seq2seq cross-entropy
Decoding/evaluation	Teacher-forced NLL; generated text metrics
Intervention stage	Inference only
Model parameters during CSI	Frozen

Table 8: Model configuration used for stream-level CSI analysis.

B Extended Discussion

Why We Use Likelihood-Based Evaluation.

Our evaluation focuses on how the model relies on different visual streams during decoding rather than on translation quality in the traditional sense. Because our claims concern internal dependency and causal contribution, we use likelihood-based measures, specifically, changes in negative log-likelihood under inference-time intervention to identify when a stream becomes necessary for predicting a token. These measures are better suited to our setting than corpus-level metrics such as BLEU, which reflect surface correspondence to reference text (Leiter et al., 2024) but are not designed to capture mechanism-level behavior (Callison-Burch et al., 2006; Vig et al., 2020; Sulem et al., 2018). This choice aligns with recent work emphasizing that explanations of model reliance should be grounded in internal evidence rather than sequence-level performance alone (Kumar et al., 2022; Vig et al., 2020). Because the body, face, and hand streams are explicit architecture-level units that can be directly suppressed, CSI measures their intervention effect rather than estimating stream importance from gradients. Gradient-based attribution therefore answers a complementary sensitivity question and, like other saliency methods, can require separate faithfulness checks (Adebayo et al., 2018; Sundararajan et al., 2017).

Why Attention Alone Falls Short as an Explanation

Attention-based analyses are widely used to interpret multimodal models because they provide an intuitive view of how information is accessed during decoding, and often align with linguistic expectations, such as increased facial attention during interrogative token generation. However, our intervention-based analyses show that attention

alone does not reliably reflect functional stream influence, supporting **H3**. Suppressing attention heads with high attention mass frequently produces negligible changes in decoding likelihood, while interventions that directly perturb stream information, such as shuffling, can induce substantial likelihood shifts even when attention signals are weak. This discrepancy indicates that attention primarily reflects where the model looks rather than which information it functionally relies on to generate predictions. Consequently, attention-based explanations are descriptive but not necessarily causal, and should be complemented with inference-time interventions when analyzing stream contributions in multistream systems.

Why not pretrained LLM/VLM backbones?

We preserve separate body, hand, and face pathways because stream identity is the intervention variable in CSI. In architectures that compress visual input into shared tokens before decoding, the same disentanglement is not readily available after fusion; applying CSI would require intervening at an identifiable pre-fusion or intermediate representation. Our late-fusion design is therefore an architectural choice for controlled analysis rather than a claim that this fusion strategy is universally preferable. Larger pretrained LLM/VLM-based translators may achieve stronger translation quality, but their intervention points would need to be defined separately.

C ASL Understanding

Signs in ASL are commonly described using a small set of linguistic parameters: handshape, palm orientation, movement, location, and non-manual markers (Stokoe, 1960; Brentari, 1998; Valli et al., 2011). These parameters work together

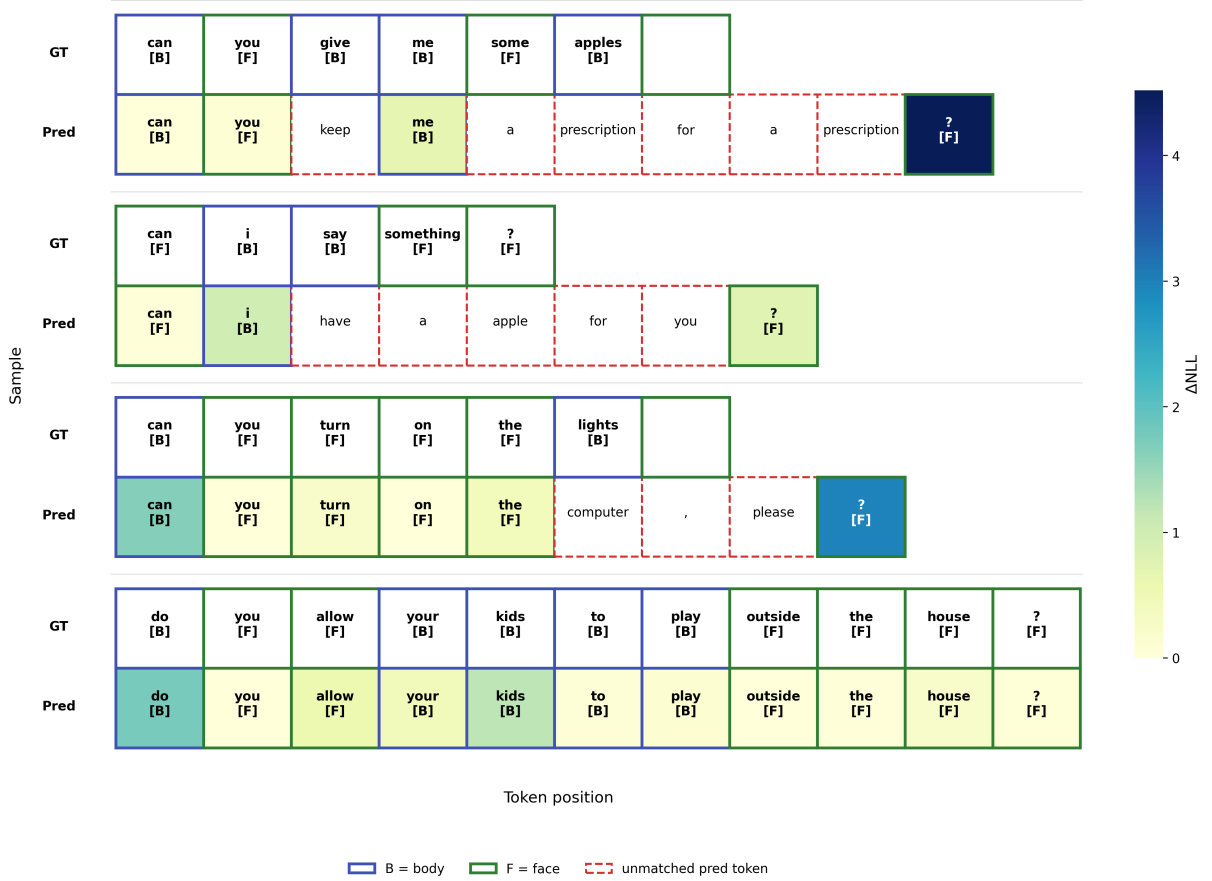


Figure 3: Most affected tokens in a Question-type sentence. Interrogative symbols like "?" cause the most disruption in the model behavior. *Examples from ASLing.*

to form meaning, and a change in one parameter can change the identity or interpretation of a sign.

In this work, we focus on the aspects that are available in our visual features. The left- and right-hand streams capture manual information related to hand configuration and movement, while the body and face streams capture non-manual cues such as posture, facial expression, and head movement. We therefore use these streams to study how manual and non-manual information contributes to SLT decoding.

Table 9 summarizes the six sentence types referenced in our analysis along with their typical non-manual markers and example glosses.

D Faithfulness

Faithfulness of raw stream signals. Our goal is to evaluate whether raw internal stream scores provide faithful evidence of causal stream importance. Let i denote a target token and $s \in \mathcal{S}$ denote a visual stream, where $\mathcal{S} = \{\text{body, left, right, face}\}$.

For a raw internal signal z , such as attention mass or router weight, we obtain a stream-specific score $z_i(s)$ before intervention.

Using intervention notation, the causal effect of ablating stream s is

$$\Delta_i(s) = NLL_i(\text{do}(\text{perturb}(s))) - NLL_i(\text{do}(\text{base})). \quad (5)$$

Here, $\Delta_i(s)$ is not itself a raw stream signal and is not directly perturbed. Rather, it is the observed causal outcome of perturbing the stream-specific model component, such as attention over the body, hand, or face stream.

We therefore evaluate whether raw stream scores are faithful by asking whether higher values of $z_i(s)$ predict larger causal effects $\Delta_i(s)$. For a signal z , the top- and bottom-ranked streams are

$$\begin{aligned} s_i^{\text{top}}(z) &= \arg \max_{s \in \mathcal{S}} z_i(s), \\ s_i^{\text{bot}}(z) &= \arg \min_{s \in \mathcal{S}} z_i(s). \end{aligned} \quad (6)$$

Table 9: ASL sentence types and associated non-manual signals. Illustrative examples use conventional ASL gloss notation: uppercase words denote lexical signs and superscript labels mark the scope/type of a non-manual signal. The glosses provide linguistic context only and are not used to assign the English-derived sentence-type labels.

Sentence Type	Non-manual Signals	Example
1. Questions		
a. Yes-No Questions	Eyebrows raised, eyes widened, head and body may be tilted forward; shoulders may be raised; last sign may be held	$\overline{y/n}$ GIRL HOME
b. Wh-Questions	Eyebrows squinted, head tilted; body may be forward; shoulders may be raised	$\overline{wh}$ GIRL WHERE
c. Rhetorical Questions	Eyebrows raised, head may be tilted or may shake slightly	$\overline{rh}$ PRO-1 TIRED WHY STUDY ALL-NIGHT
2. Negation	Head shakes side-to-side; may have frown or squint	$\overline{neg}$ GIRL HOME
3. Commands	Direct eye contact with addressee, may frown	*STAND*
4. Topicalization	Eyebrows raised, head tilted, possibly a short pause	$\overline{t}$ EXAMS, PRO-1 DETEST
5. Conditionals	Eyebrows raised, head tilted; possibly a short pause and eye gaze shift	$\overline{cond}$ TOMORROW SNOW, CLASS CANCEL
6. Declaratives	None	PRO-1 DETEST EXAMS

For compactness, we write

$$\begin{aligned}\Delta_i^{\text{top}}(z) &= \Delta_i(s_i^{\text{top}}(z)), \\ \Delta_i^{\text{bot}}(z) &= \Delta_i(s_i^{\text{bot}}(z)).\end{aligned}\quad (7)$$

Faithfulness is then measured by the top-bottom causal Diff:

$$\text{Diff}(z) = \frac{1}{N} \sum_{i=1}^N \left(\Delta_i^{\text{top}}(z) - \Delta_i^{\text{bot}}(z) \right). \quad (8)$$

A positive Diff indicates that the raw signal tends to rank causally important streams above less important streams.

We also report a win rate:

$$\text{Win}(z) = \frac{100}{N} \sum_{i=1}^N \mathbb{I} \left[\Delta_i^{\text{top}}(z) > \Delta_i^{\text{bot}}(z) \right]. \quad (9)$$

While $\text{Diff}(z)$ measures the average magnitude of the causal separation, $\text{Win}(z)$ measures how often the top-ranked stream produces a larger causal effect than the bottom-ranked stream.

Importantly, raw base NLL is not stream-specific: $\text{NLL}_i^{\text{base}}$ is a single scalar for token i and therefore cannot rank streams by itself. Similarly, $\Delta_i(s)$ is only defined after an intervention on

Sentence Type	ASLing	OpenASL	Combined
Declarative	59	590	649
Question	18	38	56
Rhetorical Question	-	127	127
Negation	4	114	118
Command	7	68	75
Conditional	2	36	38
Total	89	973	1063

Table 10: Distribution of sentence types in the ASLing and OpenASL test sets.

stream s ; it is the causal measurement, not a pre-intervention raw signal. Thus, the faithfulness analysis focuses on raw stream-specific signals such as attention scores and router weights, and evaluates them using intervention-based ΔNLL .

E Sentence-Type Annotation

The ASL datasets used in our experiments do not include sentence-type labels. Because our analysis studies whether causal stream importance differs across linguistic constructions, we assign each test example a coarse sentence-type label using the reference translation.

We use an automatic sentence-level annotator to classify each reference translation into one of

the following broad categories: question, negation, command, topicalization, conditional, or declarative. Questions include yes-no, wh-, and rhetorical questions. When the number of examples in a sub-type is small, we merge related question types for analysis. Declaratives are statement-like sentences without an explicit interrogative, negative, conditional, or command structure.

The annotator uses surface cues from the reference translation, such as question marks, interrogative words, negation markers, imperative forms, and conditional markers. For example, reference translations containing words such as “what”, “why”, “where”, “who”, or a question mark are usually labeled as questions. Sentences containing explicit negative markers such as “not”, “no”, or “never” are labeled as negation when the negative construction is central to the sentence. Examples without these markers are labeled as declaratives unless another category is more appropriate.

These sentence-type labels are not used to train the model, select model outputs, or compute the translation loss. They are used only for post-hoc grouping of CSI results. This allows us to compare whether intervention effects differ across sentence types while keeping the translation model unchanged.

To reduce annotation noise, we manually inspect a subset of the assigned labels, with particular attention to the question/declarative distinction used in the main analysis. Ambiguous examples are assigned to the closest coarse category based on the reference translation. Because the labels are derived from English reference text rather than direct linguistic annotation of the signed video, we treat the sentence-type analysis as coarse-grained rather than a complete linguistic annotation of ASL structure. The main hypothesis-driven comparisons use the question and declarative groups; lower-frequency categories are reported only as exploratory context. The post-hoc sentence-type analysis contains 89 ASLing examples, while the full ASLing test set used for baseline translation evaluation contains 90 examples.

Table 10 reports the sentence-type distribution for the ASLing and OpenASL test sets used in our post-hoc analysis. Declarative sentences form the largest category in both datasets, while questions, negations, commands, and conditionals occur less frequently. These labels are used only to group evaluation results by sentence type and are not used

Dataset	Test N	BLEU	ROUGE-L	METEOR
ASLing	90	17.04	28.9	23.58
OpenASL	973	6.951	19.76	15.87

Table 11: Baseline translation performance of the frozen SLT model before intervention. CSI analysis is performed on this trained model without updating its parameters; these values are reported to characterize the analyzed checkpoint rather than to claim state-of-the-art translation performance.

during model training.

Baseline Performance All intervention results are computed on the same frozen model reported in Table 11. This ensures that the observed Δ NLL values reflect inference-time perturbations rather than additional training or model adaptation.

Sent. type	Signal	Top-1 Δ NLL	Bottom-1 Δ NLL	Diff Δ NLL	Win (%)	<i>N</i>
Questions	Attention	0.297	0.241	0.056	44.4	18
Questions	Router	0.351	0.053	0.298	77.8	18
Questions	Δ NLL	0.656	-0.191	0.847	100.0	18
Negation	Attention	0.192	0.187	0.005	50.0	4
Negation	Router	0.192	0.192	0.000	50.0	4
Negation	Δ NLL	0.252	0.013	0.239	100.0	4
Commands	Attention	0.644	0.119	0.525	71.4	7
Commands	Router	0.644	0.249	0.395	71.4	7
Commands	Δ NLL	1.096	-0.237	1.333	100.0	7
Declarative	Attention	0.295	0.121	0.174	63.2	57
Declarative	Router	0.301	0.167	0.134	52.6	57
Declarative	Δ NLL	0.690	-0.176	0.866	100.0	57

Table 12: **Sentence-type breakdown of H3 ranked faithfulness for ASLing.** For each sentence type and ranking signal, we compare the effect of ablating the top-ranked stream against the bottom-ranked stream. The Δ NLL ranking is included only as an oracle reference and has a perfect win rate by construction; attention and router rankings are the pre-intervention signals evaluated for faithfulness.

E.1 Extended Experiments

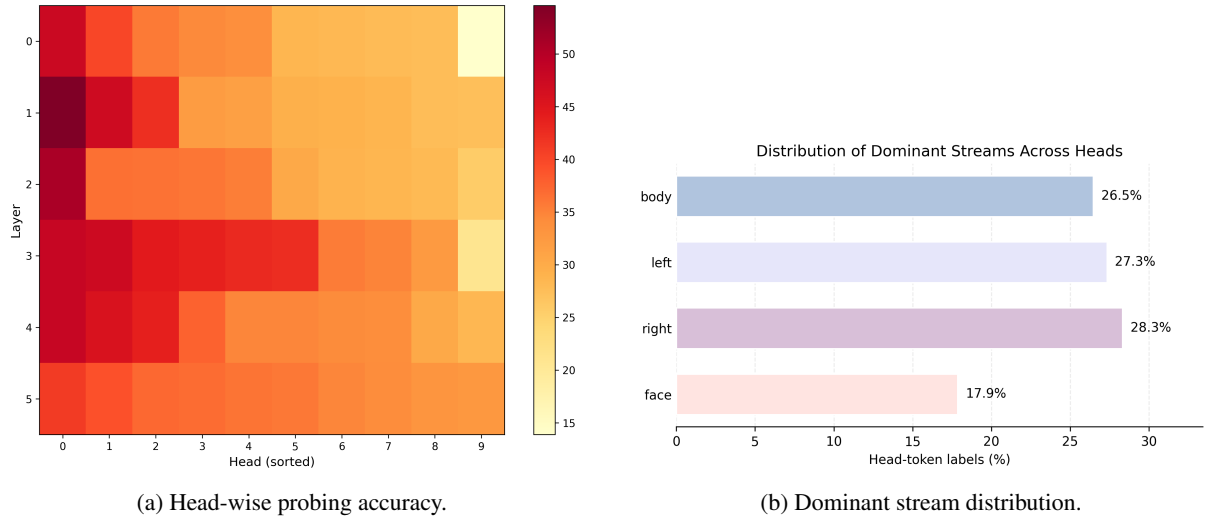


Figure 4: Head-wise probing and dominant stream distribution for OpenASL dataset based on (Li et al., 2023). The left panel shows where stream information is most decodable across heads, while the right panel shows how often each stream is selected as dominant.

Layer	Target stream	Nonmatch Δ NLL	Match Δ NLL	Diff Δ NLL
0	body	0.009	0.044	0.035
0	left	0.009	0.017	0.008
0	right	0.005	0.027	0.022
0	face	0.014	0.107	0.093
1	body	-0.007	0.044	0.052
1	left	-0.004	0.046	0.050
1	right	-0.014	0.052	0.066
1	face	0.010	0.099	0.089
2	body	-0.008	0.060	0.067
2	left	-0.009	0.106	0.115
2	right	0.009	0.115	0.106
2	face	-0.007	0.148	0.155
3	body	-0.066	0.077	0.144
3	left	-0.017	0.116	0.133
3	right	0.007	0.127	0.120
3	face	-0.032	0.140	0.172
4	body	-0.035	0.082	0.117
4	left	-0.060	0.336	0.395
4	right	0.036	0.522	0.486
4	face	-0.140	0.145	0.285
5	body	-0.065	0.314	0.379
5	left	-0.137	0.340	0.478
5	right	0.005	0.385	0.380
5	face	-0.052	0.035	0.087

Table 13: **Layer- and stream-wise causal selectivity.** For each target stream, matched interventions perturb the stream identified as dominant for the token, while nonmatched interventions perturb the remaining streams. The Diff is computed as matched Δ NLL minus non-matched Δ NLL.