Targeted Linguistic Analysis of Sign Language Models with Minimal Translation Pairs

Serpil Karabüklü[1], **Kanishka Misra**[1,2,*], **Shester Gueuwou**[1], **Diane Brentari**[3], **Greg Shakhnarovich**[1], **Karen Livescu**[1]

[1]Toyota Technological Institute at Chicago, [2]Linguistics Department, University of Texas at Austin, [3]Linguistics Department, University of Chicago

Correspondence: skarabuklu@ttic.edu

Abstract

*Models of sign language have historically lagged behind those for spoken language (text and speech). Recent work has greatly improved their performance on tasks like sign language translation and isolated sign recognition. However, it remains unclear to what extent existing models capture various linguistic phenomena of sign language, and how well they use cues from the multiple articulators used in sign language (hands, upper body, face). We introduce a new benchmark dataset for American Sign Language, ASL Minimal Translation Pairs (*ASL-MTP*), divided into multiple types of sign language phenomena and corresponding minimal pairs of translations, for performing such linguistic analyses. As a case study, we use* ASL-MTP *to analyze a state-of-the-art ASL-to-English translation model. We conduct a targeted analysis of the model by ablating various input cues during training and inference and evaluating on the phenomena in* ASL-MTP*. Our results show that, while the model performs above chance level on most of the phenomena, it relies strongly on manual cues while often missing crucial non-manual cues.*

https://github.com/serpilkarabuklu/SL-Models-Analysis

1. Introduction

Sign languages convey meaning visually through the combination of multiple articulatory channels, traditionally grouped by linguists into manuals (hands) and non-manuals (facial and body movements) [52, 53, 55, 69]. Computational models of sign language video have been developed for a variety of tasks, such as isolated sign recognition [35], continuous sign recognition (i.e. glossing) [10], and translation from sign language to written text in a spoken language [8, 61, 87]. Recent work has also developed pre-trained representation models for sign language video, in order to enable quick fine-tuning for multiple downstream tasks [15, 21, 80]. Some models of sign language have fo-

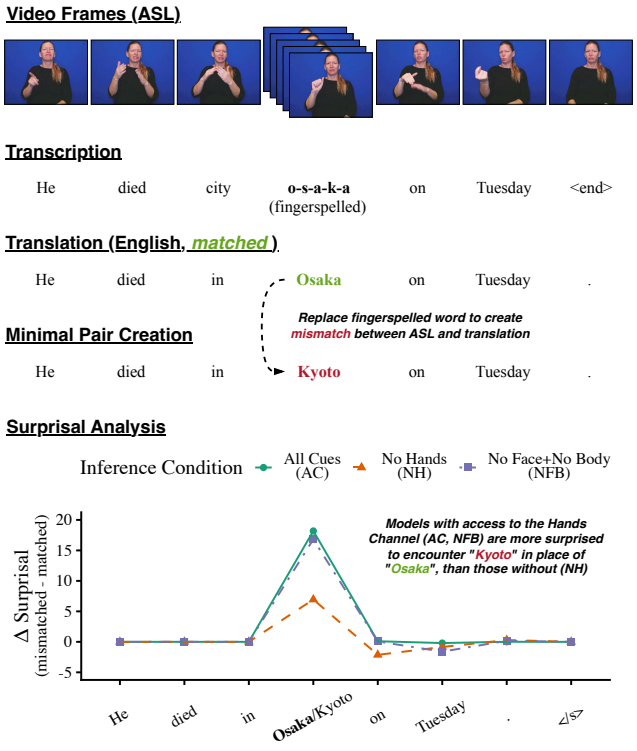

Figure 1. Our dataset construction and analysis approach. We create minimal pairs for English translations of ASL inputs by replacing critical segments (here, the fingerspelled word, "Osaka"). We then measure the difference in a sign language translation model's surprisal on the matched vs. mismatched sequences to quantify the model's sensitivity to the target phenomenon (fingerspelling). The stimulus shown is taken from FLEURS-ASL [65], and is used only for illustration.

*Work done partly while at TTIC

cused only on manual signs [19–21], some use the entire input images [54, 61], and some divide the input into multiple channels, for both perception [9, 14, 15] and generation [40, 57].

Recent work has produced substantial improvements in the performance of sign language translation models. However, it is unclear how well such models handle specific linguistic phenomena, and in particular, phenomena that critically depend on channel-specific cues. For instance, phenomena such as Wh- and Polar Questions often involve a combination of hand and eyebrow movements [3], while fingerspelling and numbers largely depend on manual cues. Even for models that explicitly learn from multiple channels, it is unclear if they extract information from these channels in a way that matches expectations from sign language linguistics.

We aim to address the abovementioned gaps, using American Sign Language (ASL) data. We first present ASL Minimal Translation Pairs (ASL-MTP), a collection of ASL utterance videos along with corresponding (matched, mismatched) pairs of English translation sentences that minimally differ with respect to a particular target phenomenon. Here, a model should assign higher probability to the correct (matched) reference translation than to its minimally differing (mismatched) counterpart. ASL-MTP covers 9 phenomena and is inspired by the now well-established practice of using minimal pair datasets to evaluate the linguistic knowledge of language models [22, 38, 42, 71, i.a.], as well as contrast sets in machine translation [59]. ASL-MTP is the first minimal-pair dataset that focuses on phenomena-specific sensitivity in sign language translation models.

As a case study, we apply minimal pair analysis with ASL-MTP to SHuBERT+ByT5 [15], a state-of-the-art sign language translation model that uses multiple input channels. This property of the model allows us to manipulate information represented in each channel, to shed light on whether the model extracts information in a manner that conforms to linguistic expectations. Specifically, we devise a set of cue ablations where one or more channels are masked from the input, and test whether doing so affects the model's performance on ASL-MTP, especially on phenomena that rely upon the ablated cue. We find that when the model has access to all cues, it performs above chance on 8 of the 9 phenomena in ASL-MTP. For one phenomenon, Polar Questions, the model has a strong bias toward declarative sentences over questions. When ablating cues, we find mixed results in terms of the effect of ablation on model performance. While the model is clearly affected by the lack of hands, it is not always sensitive to losses in non-manual cues. We also ablate cues at training time, inspired by controlled rearing of language models [e.g., 44].

Our work contributes to finer-grained linguistic evaluation of sign language translation models, and our case study points to a need to improve the use of non-manual cues in a state-of-the-art model.

2. Related Work

2.1. Sign Language Models

Sign language processing has traditionally been fragmented into task-specific methods and models, with different architectures and designs. A primary bottleneck limiting unification has been the scarcity of large datasets for pretraining. However, with the emergence of larger datasets such as YouTube-ASL [68] (∼1,000 hours), recent years have seen the rise of pretrained sign language models that can be adapted with minimal fine-tuning for a variety of downstream tasks. Both supervised [68, 87] and self-supervised approaches [54, 80] have advanced considerably, but most large-scale ASL models remain inaccessible as their weights are not publicly released. The only state-of-the-art ASL model of which we are aware that is publicly available and trained on substantial ASL data is SHUBERT [15], which we use for our case study here. SHUBERT is a self-supervised ASL video representation model, trained via masked prediction of automatically learned discrete "tokens" in multiple channels (hands, face, upper body pose), and has been fine-tuned to obtain state-of-the-art performance on multiple tasks (translation, isolated sign recognition, fingerspelling detection). SHUBERT is described in more detail in Sec. 4.1.

Evaluation of sign language models has typically relied on high-level, task-specific measures such as BLEU [49] and BLEURT [58] for translation [8, 61] and accuracy for isolated sign recognition [31, 35, 46]. Some studies have focused on isolated signs at the phonological level [30, 32, 56, 67], while others on continuous signing have evaluated models on individual linguistic phenomena, such as intensification [24], phonetic reduction driven by discourse effects [23], and co-reference resolution of indexical signs [85]. Task-specific measures often miss out on linguistic nuances of sign language, while the existing targeted analyses are limited to a single channel (the hands) or a constrained setting (e.g., isolated signs).

2.2. Linguistic Analysis of Language Models

In contrast with standard task-oriented benchmarks, linguistically motivated analyses of language models (LMs) typically involve a controlled approach to collecting data that isolates a phenomenon of interest [17, 38, 43, 74, 79]. A standard approach to evaluation in this space is the minimal pairs paradigm [22, 71, etc.], in which the LM is provided with pairs of sentences that differ in ways that are critical to the phenomenon of interest, where one of them is acceptable and the other unacceptable. As an example of a pair that can be used to evaluate an LM on number agreement, consider "*The*

*woman **laughs**.*" vs. "**The woman **laugh**.*", which differ only in the number agreement of the verb (*laugh*) with the subject (*woman*). The LM is evaluated on a number of such pairs, by comparing its "score" (usually, log-probability per token) on each pair. Accuracy, then, is the proportion of time the LM's score for the acceptable sentence is greater than for the unacceptable sentence.

There has been a concerted effort to build controlled minimal-pair benchmarks for languages beyond written English [26, 62–64, 81]. The idea of *translation* minimal pairs has been used to study machine translation models' handling of specific linguistic phenomena like agreement and polarity [59]. ASL-MTP takes inspiration from these and expands the tradition of minimal pair analysis to sign languages.

Recent studies have also analyzed linguistic behavior in LMs by performing controlled training data ablation [25, 34, 44, 50, 82, 84]. In these studies, targeted parts of an LM's training corpus are removed or "ablated" to test whether models can recover this knowledge from other parts of the corpus. In our case study, we take (loose) inspiration from this approach and ablate specific channels in the model (hands, body, face).

2.3. Manual and non-manual channels

Our work targets phenomena encoded across multiple channels, including the hands, face, and body. Signs encoded in the hands are referred to as manual signs while facial and body movements that convey grammatical functions are referred to as non-manuals [28, 55, 69, i.a.]. For some phenomena, such as Fingerspelling [7, 29] or Classifiers [4, 88], the primary cues are only in the hands (manual signs); others, such as Wh-Questions [3, 45], Negation [45, 70], and Conditionals [2, 36, 37, 76, 78], have primary cues which are both manual signs and non-manuals. For instance, Conditionals are conveyed with both the manual sign IF and the non-manual eyebrow raise. Finally, some sign language phenomena are solely encoded through non-manuals. For instance, a Polar Question (*Are you ready?*) differs from its declarative counterpart (*You are ready.*) only in terms of eyebrow raise [3, 73].

In general, manuals are considered to convey the bulk of the content in sign language [6], but other cues also contribute substantially, either as primary or secondary cues [5, 41]. In fact, non-manual cues are prevalent across linguistic domains [51, 77]: phonology [75], morphology [1], syntax [36, 37, 45, 72], semantics and pragmatics [11, 18, 27, 60]. We therefore expect successful sign language models to use information from both manual and non-manual channels.

3. ASL Minimal Translation Pairs

One of the main contributions of this work is ASL Minimal Translation Pairs (ASL-MTP), a dataset to evaluate fine-grained linguistic capacities of models that translate ASL to English. ASL-MTP consists of 1,275 ASL videos along with corresponding pairs of acceptable and unacceptable written English translations. Since the acceptability of a sentence depends on the extent to which it matches the ASL video, we call the acceptable sentences "matched" and unacceptable ones "mismatched". The dataset is divided into 9 subsets, each of which targets a specific phenomenon.

The ASL videos and their corresponding sentences were drawn from ASLLRP [47], a collection of 2,048 high-quality, linguistically annotated ASL utterances, along with their English translations, produced by 4 signers. ASLLRP includes annotations for manual signs (e.g., number of hands, hand movements) and time-aligned non-manuals (e.g., head position and movements, mouth movements, eye gaze) along with their grammatical functions (e.g., classifier, question, conditional). These annotations, combined with the fact that ASLLRP has not been widely used in the training of sign language models, make it an ideal source for evaluating models. Although the dataset is not large, it provides enough data to evaluate models on the phenomena of interest and, as we will see, to obtain statistically significant results in our analyses (Sec. 4). Below we provide more details about our phenomenon selection criteria and dataset construction, as well as the intended usage of ASL-MTP for minimal pair evaluation.

3.1. Details of ASL-MTP construction

Phenomena To investigate whether sign language models rely on cues from multiple input channels, we selected 9 phenomena that involve a range of channel combinations. Our phenomena can be grouped into three subsets: 1) ones that are mainly encoded in the hands—Numbers, Fingerspelling, and Classifiers; 2) ones that are encoded in both the hands and face—Negation, Wh-Questions, and Conditionals; and 3) ones that are predominantly encoded in the face—Polar Questions. This grouping is not perfectly clean, because of the existence of varying secondary cues—e.g., there are several stimuli in our dataset where 'Conditionals' are signed using non-manual cues. Therefore, we will discuss results on such exceptional cases separately, when relevant.

Dataset Construction Tab. 1 shows a detailed description of our stimuli design methods, across the 9 phenomena. Our general stimuli construction is as follows. First, we queried ASLLRP for instances (consisting of a video, its glossed version, and its English representation) suited for a given phenomenon. Then, for each instance, we manipulated its English translation by replacing certain words or rewrit-

Phenomenon	Description	Construction	Minimal Pair Examples
Numbers (N=119)	Manual signs with phonetically complex handshapes and movements.	**Filtering Criteria:** Sentences containing numerical values. **Manipulation:** Replace number with a different number.	**Matched**: The movie starts at *7*. **Mismatched**: The movie starts at *8*.
Fingerspelling (N=170)	Sequences of letter representations of spoken words, signed using handshapes. Mostly used for proper nouns, technical terms, and borrowings without established signs.	**Filtering Criteria:** Gloss annotations for fingerspelling (e.g. '#A-N-N'). **Manipulation:** Replace fingerspelled word with another contextually acceptable word.	**Matched**: *Ann* hates fish but likes chicken. **Mismatched**: *Beth* hates fish but likes chicken.
Classifiers (N=150)	Signs denoting the salient semantic properties of entities like size, shape, and number. Handshape tends to refer to entities and movement to events.	**Filtering Criteria:** Gloss annotations with "cl:" in their prefix. **Manipulation:** Replace word signed as a classifier with another contextually acceptable word.	**Matched**: Are *the friends (with a two-handed DCL:C†)* going out? **Mismatched**: Is *the friend* going out?
Conditional Statements (N=205)	*If..., then...* statements. Conveyed with the manual sign IF with eyebrow raise.	**Filtering Criteria:** Sentences containing the conditional marker "if". **Manipulation:** Replace *if* with *when*.	**Matched**: *If* I see my friend, I will be thrilled. **Mismatched**: *When* I see my friend, I will be thrilled.
Negation vs. Positive (N=104)	Expressions of negative polarity in a sentence, often through the manual sign NOT and the non-manual headshake.	**Filtering Criteria:** Instances containing explicit negation markers (i.e. "not" or contractions with "n't"). **Manipulation:** Remove negation.	**Matched**: Bob *hasn't* sent the letter. **Mismatched**: Bob *has* sent the letter.
Positive vs. Negation (N=104)	Expressions of positive polarity in a sentence, without negative manual signs or headshake.	**Filtering Criteria:** Instances not containing explicit negation markers (i.e. "not" or contractions with "n't"). **Manipulation:** Add negation word.	**Matched**: Bob *read* a book. **Mismatched**: Bob *didn't read* a book.
Wh-Questions (N=123)	Content questions formed by wh-signs (what, who, where, when, why, how) and the non-manuals eyebrow lowering and/or head tilt.	**Filtering Criteria:** Sentences ending with a "?" and containing at least one wh-word ("what", "when", "who", "where", "why", "whom", or "how"). **Manipulation:** Replace wh-word with another acceptable wh-word.	**Matched**: *When* did father arrive home? **Mismatched**: *How* did father arrive home?
Polar Questions vs. Declaratives (N=150)	Polar questions formed only by the non-manual eyebrow raise.	**Filtering Criteria:** Sentences ending with a "?", followed by manual verification for Polar Questions. **Manipulation:** Convert to declarative.	**Matched**: *Are Jen and Joe married?* **Mismatched**: *Jen and Joe are married.*
Declaratives vs. Polar Questions (N=150)	Declarative sentences which do not involve any question.	**Filtering Criteria:** Sentences that do not end with a "?". **Manipulation:** Convert to polar question.	**Matched**: *All the men left together.* **Mismatched**: *Did all the men leave together?*

Table 1. Phenomena included in ASL-MTP, along with their sample sizes, descriptions, construction, and examples. † In the example for classifiers, we show the difference in classifiers using ASLLRP notation: DCL - descriptive classifier, C - handshape. This classifier refers to a group of friends and cannot refer to a singular entity.

ing it to target the phenomenon in question. Taking "Polar Questions vs. Declaratives" as an example, we rewrote the matched sentence *Are Jen and Joe married?* in its declarative form to create its mismatched counterpart: *Jen and Joe are married.* Importantly, both the matched and mismatched utterances are grammatically correct—they differ in whether they are a correct translation of the input ASL video, and specifically in terms of the phenomenon in question. All of these considerations, applied to ASLLRP, yield the focused dataset ASL-MTP of 1,275 pairs across phenomena[1].

3.2. Using ASL-MTP for Sign-Conditioned Minimal Pair Analysis

To analyze a model's behavior on the linguistic phenomena described above, we adopt standard practice in minimal-pair evaluation (see Sec. 2.2 for an overview), and compare the model's log-probabilities on the sentences in each pair, when conditioned on the ASL input. Let $\mathcal{D}$ =

[1] ASL-MTP can be found here: https://github.com/serpilkarabuklu/SL-Models-Analysis/blob/main/data/asl-mtp.csv

$\{(F_1, a_1, u_1), \ldots, (F_n, a_n, u_n)\}$ be a phenomenon-specific dataset whose entries comprise input features extracted from the sign language video $F_i \in \mathbb{R}^{T \times d}$ (where T is the number of frames) involving the phenomenon, a matched, ground-truth reference sentence translation of the video a_i, and a minimally differing sentence u_i that has been perturbed in a targeted manner to be mismatched. We expect a model that has mastery over the target phenomenon to find the mismatched sentence u_i more 'surprising', or unlikely, than the matched sentence a_i, when conditioned on the video F_i. We measure the model's (un)likelihood for a sentence $s_i := (x_1, \ldots, x_{|s_i|})$ by computing its conditional, per-token surprisal (negative log-probability):

$$\mathcal{S}(s_i) = \frac{1}{|s_i|} \sum_{t=1}^{|s_i|} -\log p(x_t \mid x_{<t}, F_i) \tag{1}$$

We then compute the difference in surprisals for the mismatched and matched sentences:

$$\Delta\text{Surprisal}_i = \mathcal{S}(u_i) - \mathcal{S}(a_i) \tag{2}$$

Insofar as a model is sensitive to the phenomenon that governs the differences between a_i and u_i, we expect $\Delta\text{Surprisal}_i$ to be greater than 0. Accuracy, then, is the proportion of pairs for which $\Delta\text{Surprisal} > 0$. Since this comparison is done over pairs, chance performance is 50%.

The basic use case we envision for ASL-MTP, then, is to evaluate accuracy of an ASL-to-English translation model (in the sense of accuracy defined above) on the 9 phenomena-specific subsets. This framework provides a general method to evaluate a model on a number of phenomena for which minimal pairs can be created, given a fixed video. In our own case study (Sec. 4), we further divide 2 of the 9 phenomena into two subsets each, corresponding to those examples that rely on non-manuals only vs. both manuals and non-manuals.

4. A Case Study

Next we use ASL-MTP for a case study, in which we analyze an open, state-of-the-art ASL-to-English translation model. Specifically, we use ASL-MTP to (1) study the model's behavior on the 9 phenomena, in terms of the surprisal-based accuracies defined above, and (2) analyze the extent to which the model uses cues from the hand, body, and facial information channels.

4.1. Model studied

Based on our research goal above, we define a set of criteria that govern our choice of model: First, the model must take in ASL video input and produce English translations, in a manner that allows the extraction of token probabilities (to facilitate analysis via surprisals). Second, it should enable control over the input channels that encode information from the cues we aim to study—e.g., one should be able to mask out information from the eyes while preserving other cues (hands, mouth, body movements).[2] Finally, the model weights, training data, and training pipeline should be openly available, and runnable on academic compute.

The only currently available model that meets these criteria is the SHUBERT+ByT5 translation model from Gueuwou et al. [15].[3] This translation model combines two jointly fine-tuned models: SHUBERT, a BERT-style [12] encoder pretrained on 1,000 hours of continuous ASL YouTube videos (combining subsets of YouTube-ASL [68] and YouTube-SL-25 [66]), and the ByT5-Base text translation model [83].

SHUBERT takes inputs that are decomposed into four channels: face (mouth and eye image crops), left hand crops, right hand crops, and body pose keypoints. The face and hand crops are represented using DINOv2 image features [48]. The input channels can be manipulated in order to measure SHUBERT's use of information restricted to a particular channel. The translation model is trained to map from ASL videos to English translations, using the next-token prediction objective (here, the tokens are bytes) in an autoregressive manner. This means that we can use its next-token probabilities to compute the log-probabilities needed for our minimal pair analysis (see Sec. 3.2) In all of our experiments, we either (1) use the translation model as is while manipulating the input cues (i.e., at inference time) or (2) re-train versions of SHUBERT with varying cues present, which we again jointly fine-tune with ByT5, directly following the training pipeline (including hyperparameters) of Gueuwou et al. [15].

4.2. Cue Ablation

To understand the importance of different visual cues to the model's performance, we use a systematic ablation strategy that masks specific features (corresponding to targeted cues) in the input video frames. For example, if masking the hands does not affect performance on a particular phenomenon, this indicates that the model is not sensitive to handshape and orientation (i.e., does not use these cues in predicting the token probabilities) in examples of that phenomenon.[4] We use the keypoints returned from the MediaPipe library [39] to detect the regions of interest, which are then selectively greyed out in the video frames (see Fig. 2). We perform these ablations either at inference time (Sec. 4.3) or during training (Sec. 4.5).

We use the original SHUBERT+ByT5 model with the full video input as our baseline condition, where no masking

[2] We note that this criterion is not necessary to use ASL-MTP or to do surprisal-based analysis, but only to carry out our case study which concerns the analysis of cue use.

[3] `https://shubert.pals.ttic.edu/`

[4] In this case, the model may still be sensitive to hand *location*, which is encoded in the body pose.

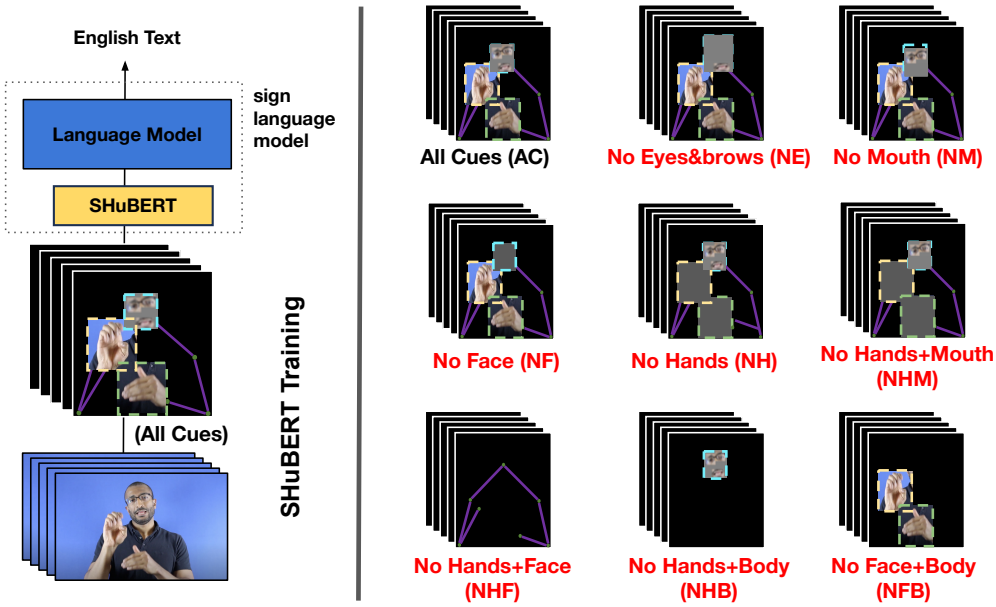

Figure 2. **Left:** A depiction of how SHUBERT [15] is combined with an off-the-shelf language model (here, ByT5) to perform ASL-to-English translation. **Right:** Examples of inputs provided to the model for the All Cues condition as well as the 8 Cue Ablations.

is performed, and refer to this as **All Cues (AC)**. Then, to isolate the impact of specific cues, we consider 8 ablations: 1) **No Eyes & Brows (NE)**, where the eye and eyebrow regions (involved in questions [3] and conditionals [36, 78]) are masked from the face channel; 2) **No Mouth (NM)**, where the mouth region is masked from the face channel, thereby removing mouthing cues, which are often used for disambiguation, or convey adjectival or adverbial meanings; 3) **No Face (NF)**, where the entire face channel (eyes, eyebrows, and mouth) is masked; 4) **No Hands (NH)**, where the hand channels are masked; 5) **No Hands & Mouth (NHM)**, where the mouth region of the face channel and the hand channels are masked; 6) **No Hands & Face (NHF)**, where the hand and face channels are masked, thereby allowing us to test if the model can use body pose alone, which is used to indicate role shifting, contrast, and spatial organization; 7) **No Hands & Body (NHB)**, where only the face channel is retained; and 8) **No Face & Body (NFB)**, where only the hand channels are retained.

4.3. Experiment 1: Effect of Cue Ablations on Minimal Translation Pair Performance

Our first experiment evaluates SHUBERT+ByT5 on ASL-MTP, specifically focusing on the effect of ablating input cues at inference time, as discussed in Sec. 4.2. We quantify the extent to which a model is sensitive to a given set of cues by comparing its performance when those cues are ablated to performance in the All Cues (AC) condition. Tab. 2 shows the phenomenon-specific accuracies obtained by SHUBERT+ByT5, each corresponding to a particular cue ablation. We split the 'Conditionals' and 'Polar Questions vs. Declaratives' subsets of ASL-MTP into two rows each—one corresponding to videos where the phenomenon is represented *only* using non-manual cues, and the other where both manual and non-manual cues are used, giving us a total of 11 subsets. We do this to enable finer-grained analysis of model performance on cases exclusively requiring sensitivity to non-manuals. Fig. 3 shows average ΔSurprisal values across all phenomena and channel ablations.

Results with all cues We first summarize the results in the "All Cues" condition, as this serves as the baseline against which we will compare subsequent cue ablation results. We find that the model performs above chance (50%) on 9 out of 11 subsets, showing particularly good performance on Numbers, Fingerspelling, Wh-Questions, and Negation vs. Positive, while performing substantially below chance on both "Polar Questions vs. Declaratives" subsets (we return to this below). There is no clear relationship between performance and the primary or secondary cue(s) involved in a phenomenon, nor is there one between performance and the cardinality of the categories involved; for example, Numbers and Fingerspelling have large vocabularies while Negation vs. Positive is a binary distinction, and all of these have among the highest performance. Among the phenom-

ena on which the model performs above chance, it performs the worst on Classifiers. Classifier meanings critically rely on referents within the utterance [16, 88], suggesting that analyses of classifier sign errors and reference may be a good direction for future work.

Heavy reliance on hands When the hands are ablated (i.e., in the NH, NHM, NHF, and NHB conditions), the model performs significantly worse than in the AC condition (often worse than or close to chance) on Numbers, Fingerspelling, Classifiers, Wh-Questions, Positive vs. Negation, and Conditionals (NM only). This result is expected, as hands are the most important source of information in sign language in general [41] as well as a primary cue for these phenomena.[5] When hands are not a primary cue (e.g. for the Conditionals (NM only) subset), there is no significant performance reduction when ablating the hands.

Poor sensitivity to non-manual cues The model is much less sensitive to non-manual cues (e.g., head movements, eyebrow raises), even on phenomena that explicitly rely on these cues—Wh-Questions, Negation vs. Positive, Conditionals, and Polar Questions vs. Declaratives. That is, its accuracies on these phenomena are no different from those in the AC condition when these cues are ablated. For example, on the subset of the Conditionals that *exclusively* rely on non-manual cues, we notice model insensitivity across *all* cue-ablation conditions, but this could also be due to the relatively small sample size (50). The only cases where we do observe sensitivity to non-manual cues are in phenomena that do not necessarily rely on them—e.g., the model is significantly worse (relative to AC) in the absence of the face and body pose (NF and/or NFB) for Numbers, Fingerspelling, Classifiers, and Positive vs. Negation. This could be because mouthing is sometimes used to disambiguate certain signs, even when this is not a necessary cue, and the presence of mouthing is not annotated in the data. Overall, the model is not as sensitive to critical non-manual cues as we might expect from linguistic intuition, despite having access to these cues during training.

Declarative bias Among the phenomena tested, there was a particularly wide gap between model accuracies on "Declaratives vs. Polar Questions" and "Polar Questions vs. Declaratives" in *all* experimental conditions. In particular, the model shows a bias towards generating declarative sentences over polar questions, in all cue ablation settings (seen in both Table 2 and the ΔSurprisal results in Fig. 3.

[5]When we remove hands but not body pose (NH, NHM, NHF), we retain information about hand location from the body pose, but lose important handshape/orientation features.

4.4. Experiment 2: Surprisals vs. BLEURT

The results thus far suggest that SHUBERT+ByT5 is not always sensitive to the various cues it is trained to use. While we base these findings on our proposed surprisal analysis, to what extent could they also be explained using standard machine translation (MT) metrics, like BLEURT [58]? That is, what does the minimal pair analysis buy us above and beyond off-the-shelf translation measures?

An a priori argument against BLEURT (and other general MT metrics) is that it is a global similarity measure between reference and translation, and is not guaranteed to be systematically sensitive towards the phenomena in ASL-MTP. Taking Wh-Questions as an example, a model might succeed at recognizing the right Wh-word but produce a completely wrong translation: In one example (in the All Cues condition), where the reference is ***Why*** *do you have to move out of San Diego?*, the model produces ***Why*** *do you think this is ASL?*. Here, the BLEURT score is poor (24.4) but it is not due to an insensitivity to the Wh-word, but instead due to the completely different meanings encoded in the two sentences because of other word substitutions.

To confirm our a priori intuition, we obtain hypothesized translations of the ASL-MTP inputs from the SHUBERT+ByT5 model (using beam search, as in Gueuwou et al. [15]), and compute the average BLEURT scores between the model translations and the ground-truth reference sentences. We report the resulting BLEURT scores across phenomena and cue ablations in Tab. 3.

These results indeed confirm that BLEURT cannot uncover the distinctions in model behavior across phenomena and input conditions that we found in the minimal pair analysis. First, there is very little variability in BLEURT scores across phenomena. Presumably, as in our example above, BLEURT is dominated by various differences in translations besides the specific ones we target. Second, for most phenomena BLEURT is lower whenever the hands or body are removed, again with little distinction among phenomena. BLEURT is therefore unable to discover the model's insensitivity to certain non-manual cues in some phenomena. Finally, we also measure the Pearson correlation between BLEURT and surprisal-based accuracy for each phenomenon, and find generally poor to moderate correlations (ranging from -.17 for 'Polar Questions vs. Declaratives' to .36 for Numbers). These results are not surprising, but reinforce the role of minimal pair analysis, which can help diagnose specific, linguistically interpretable model behaviors that translation metrics do not reveal.

4.5. Experiment 3: Controlled Rearing of SHUBERT Using Cue Ablations

There are multiple possible explanations for the results of the inference-time cue ablation analysis of Experiment 1. One possibility is that the ablated inputs are out of distribution

Phenomenon	Primary Cue(s)	Secondary Cue	#	AC	NE	NM	NF	NFB	NH	NHM	NHF	NHB
1. Numbers	Hands		119	0.87	0.82	0.78	**0.77**	**0.75**	**0.61**	**0.58**	**0.55**	**0.55**
2. Fingerspelling	Hands		170	0.78	0.75	**0.70**	0.72	**0.68**	**0.49**	**0.46**	**0.43**	**0.49**
3. Classifiers	Hands		150	0.63	0.62	0.59	0.59	**0.53**	**0.51**	**0.48**	**0.49**	**0.47**
4. Wh-Questions	Hands + Brow lowered	Head shake	123	0.75	0.74	0.76	0.76	0.72	**0.66**	**0.65**	0.67	**0.64**
5. Negation vs. Positive	Hands + Head shake		104	0.80	0.77	0.77	0.77	0.71	0.69	0.69	0.68	0.69
6. Positive vs. Negation	Hands		104	0.65	0.68	0.62	0.66	**0.53**	**0.34**	**0.34**	**0.28**	**0.29**
7a. Conditionals	Hands + Brow raise	Head thrust, Body forward	155	0.70	0.72	0.66	0.70	0.72	**0.56**	**0.51**	**0.54**	**0.59**
7b. Conditionals (*NM only*)	Brow raise	Head thrust, Body forward	50	0.68	0.68	0.68	0.70	0.76	0.60	0.62	0.60	0.64
8. Declaratives vs. Polar Questions	Hands		150	0.97	0.97	0.97	0.96	0.96	0.96	0.95	0.95	0.95
9a. Polar Qs vs. Declaratives	Hands + Brow raise	Head forward	57	0.04	0.04	0.04	0.05	0.04	0.05	0.05	0.04	0.04
9b. Polar Qs vs. Declaratives (*NM only*)	Brow raise	Head forward	93	0.09	0.09	0.09	0.09	0.10	0.15	0.14	0.09	0.13

Table 2. Phenomenon-wise surprisal-derived accuracies across **inference conditions**. Accuracy values are **boldfaced** if they are significantly different from the accuracy on the 'All Cues' (AC) inference condition ($p < .05$, as measured by a two-tailed exact binomial test, with the Bonferroni correction for multiple comparisons). "(*NM only*)" indicates that the stimuli in that subset involve only non-manual cues. "Hands" refers to any number of cues related to handshape and orientation. Chance performance is 50%.

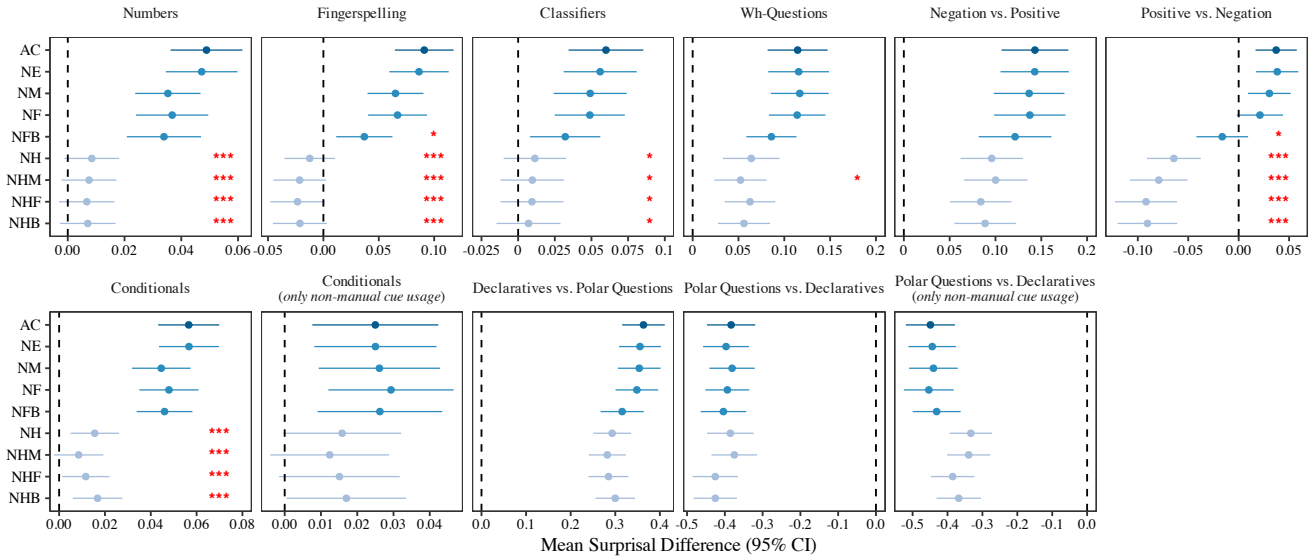

Figure 3. Average difference in surprisal of mismatched and matched sentences across phenomena and across **inference cue ablations**. Error bars indicate 95% confidence intervals. Stars (*) indicate significance test results for comparing the surprisal difference in a given cue ablation to surprisal difference in the 'All Cues' condition (AC). *: $p < .05$; **: $p < .01$; ***: $p < .001$

Phenomenon	Primary Cue(s)	Secondary Cue	#	AC	NE	NM	NF	NFB	NH	NHM	NHF	NHB
1. Numbers	Hands		119	0.50	0.50	0.47	0.48	**0.43**	**0.28**	**0.28**	**0.26**	**0.32**
2. Fingerspelling	Hands		170	0.43	0.42	**0.40**	0.41	**0.35**	**0.29**	**0.26**	**0.26**	**0.29**
3. Classifiers	Hands		150	0.44	0.43	**0.41**	**0.41**	**0.36**	**0.28**	**0.26**	**0.26**	**0.29**
4. Wh-Questions	Hands + Brow lower	Head shake	123	0.54	0.53	0.53	0.55	**0.44**	**0.30**	**0.23**	**0.26**	**0.28**
5. Negation vs. Positive	Hands + Head shake		104	0.51	0.50	**0.45**	0.47	**0.40**	**0.26**	**0.25**	**0.25**	**0.29**
6. Positive vs. Negation	Hands		104	0.47	0.47	0.44	0.45	**0.39**	**0.29**	**0.26**	**0.24**	**0.29**
7a. Conditionals	Hands + Brow raise	Head thrust, Body forward	155	0.45	0.45	0.43	0.43	**0.38**	**0.28**	**0.25**	**0.27**	**0.31**
7b. Conditionals (*NM only*)	Brow raise	Head thrust, Body forward	50	0.41	0.40	0.41	0.39	0.37	**0.26**	**0.25**	**0.26**	**0.27**
8. Declaratives vs. Polar Qs	Hands		150	0.51	0.50	0.48	0.48	**0.40**	**0.27**	**0.24**	**0.24**	**0.29**
9a. Polar Qs vs. Declaratives	Hands + Brow raise	Head forward	57	0.41	0.41	0.39	0.40	0.36	**0.21**	**0.21**	**0.23**	**0.23**
9b. Polar Qs vs. Declaratives (*NM only*)	Brow raise	Head forward	93	0.52	0.50	0.48	0.49	**0.45**	**0.20**	**0.17**	**0.25**	**0.25**

Table 3. Phenomenon-wise BLEURT scores across **inference conditions**. Values are **boldfaced** if they are significantly different from the BLEURT score in the 'All Cues' (AC) inference condition ($p < .05$, as measured by a t-test, with the Bonferroni correction for multiple comparisons). "(*NM only*)" indicates that the stimuli in that subset involve only non-manual cues. "Hands" refers to any number of cues related to handshape and orientation.

Phenomenon	Primary Cue(s)	Secondary Cue	#	AC	NF	NFB
1. Numbers	Hands		119	0.87	**0.76**	**0.73**
2. Fingerspelling	Hands		170	0.78	0.74	**0.64**
3. Classifiers	Hands		150	0.63	0.63	0.58
4. Wh-Questions	Hands + Brow lower	Head shake	123	0.75	**0.54**	**0.63**
5. Negation vs. Positive	Hands + Head shake		104	0.80	0.81	0.76
6. Positive vs. Negation	Hands		104	0.65	0.62	0.62
7a. Conditionals	Hands + Brow raise	Head thrust, Body forward	155	0.70	**0.89**	0.61
7b. Conditionals *(NM Only)*	Brow raise	Head thrust, Body forward	50	0.68	0.72	0.46
8. Declaratives vs. Polar Questions	Hands		150	0.97	0.96	0.95
9a. Polar Qs vs. Declaratives	Hands + Brow raise		57	0.04	0.09	0.07
9b. Polar Qs vs. Declaratives *(NM Only)*	Brow raise	Head forward	93	0.09	0.11	0.04

Table 4. Phenomenon-wise accuracies across **training conditions**. The input channels match the training condition in each column. Values are **boldfaced** if they are significantly different from the accuracy in the 'All Cues' (AC) inference condition ($p < .05$, as measured by a two-tailed exact binomial test, with the Bonferroni correction for multiple comparisons). "(*NM only*)" indicates that the stimuli in that subset only involved the usage of non-manual cues. "Hands" refers to any number of cues related to handshape and orientation. Chance performance is 50%.

for the model, since it is trained on the full set of cues, so we may not know how the model would behave if it were both trained and tested with the ablated input. In particular, one may wonder whether decreased performance when cues are ablated is due to the train-test mismatch and not due to the missing cue information.

To address this issue, we run "controlled rearing" experiments [25, 33, 44, a.o.], where we train new variants of SHUBERT with certain channels removed during training: We mask out the features of the ablated channel(s), re-train SHUBERT using the masked input, then combine it with ByT5 and fine-tune as for the original SHUBERT+ByT5. We follow the fine-tuning recipe in Gueuwou et al. [15]: We first fine-tune on a large corpus of weakly aligned ASL-English pairs (∼800K samples from the union of YouTube-ASL [68] and the ASL part of YouTube-SL-25 [66]), and then continue fine-tuning on a smaller (∼200K samples) but more accurately aligned training set consisting of How2Sign [13], ASL Stem Wiki [86], and OpenASL [61].

We conduct this experiment for two types of channel ablations: One where the face is removed (NF), and one where only the hands are retained (NFB). We perform our surprisal-based minimal pair analysis, and compare accuracies in these conditions to those in the AC condition. Tab. 4 shows the phenomenon-specific accuracies, while Fig. 4 in Appendix A shows average ΔSurprisal values.

For most phenomena, we see similar relative changes from the AC condition, and the models are still susceptible to declarative bias. While there are some differences (notably for Wh-Questions and Conditionals), there is no consistent improvement in performance between the inference-time ablations and the controlled rearing setting. This suggests that our inference-time ablation results are not explained away by train-test mismatch. Additional investigation of the reasons behind differences across phenomena is left for future work.

5. Conclusion

ASL-MTP is, to our knowledge, the first dataset for minimal pair analysis of linguistic phenomena in sign language translation models. We designed ASL-MTP to include various sign language structures that use distinct channels (hands, face, or body) to convey information. As a case study, we have used ASL-MTP to analyze the strengths and weaknesses of a state-of-the-art ASL-to-English translation model. We find that the model performs at above chance on almost all tested phenomena, and in cue ablation studies shows strong sensitivity to its hand input channel, but inconsistent sensitivity to non-manual channels, despite being trained on multi-channel inputs. Lastly, we have confirmed that standard machine translation evaluation (namely, BLEURT) cannot uncover the same detailed distinctions as minimal-pair analysis using ASL-MTP. Overall, the minimal-pair analysis captures nuanced distinctions across linguistic structures and specific errors, helping pinpoint targeted improvements for future sign language models.

We hope that our work will inspire additional linguistic studies of sign language models using ASL-MTP, as well as additional minimal pair benchmarks, adding to the tradition of linguistic analyses of language models. To the extent that more sign language translation models will be released publicly, ASL-MTP will enable comparative studies across models. Other potential directions for future work include building larger datasets, perhaps via automatic or semi-automatic discovery of phenomena-specific subsets in sign language video corpora, and extension to additional languages.

Acknowledgments

Kanishka Misra is supported by the Donald D. Harrington Faculty Fellowship at UT Austin.

References

[1] Diane E Anderson and Judy Reilly. 1998. PAH! The acquisition of adverbials in ASL. *Sign Language & Linguistics*, 1(2):117–142.

[2] Charlotte Lee Baker and Carol Padden. 1978. Focusing on the nonmanual components of ASL. In Patricia Siple, editor, *Understanding Language Through Sign Language Research,*, pages 27–57. Academic Press.

[3] Charlotte Lee Baker-Shenk. 1983. *A Microanalysis of the Nonmanual Components of Questions in American Sign Language*. Ph.D. thesis, University of California, Berkeley.

[4] Elena Benedicto and Diane Brentari. 2004. Where did all the arguments go?: Argument-changing properties of classifiers in ASL. *Natural Language & Linguistic Theory*, 22(4):743–810.

[5] C Fabian Benitez-Quiroz, Kadir Gökgöz, Ronnie B. Wilbur, and Aleix M Martinez. 2014. Discriminant features and temporal structure of nonmanuals in American Sign Language. *PloS ONE*, 9(2):e86268.

[6] Diane Brentari. 2019. *Sign Language Phonology*. Cambridge University Press.

[7] Diane Brentari and Carol A Padden. 2001. Native and foreign vocabulary in American Sign Language: A lexicon with multiple origins. In Diane Brentari, editor, *Foreign Vocabulary in Sign Languages*, pages 87–119. Psychology Press.

[8] Necati Cihan Camgöz, 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, CVPR 2018*, pages 7784–7793, Salt Lake City, Utah, USA.

[9] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision–ECCV 2020 Workshops Proceedings, Part IV 16*, pages 301–319, Glasgow, UK.

[10] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

[11] Geoffrey Coulter. 1978. Raised eyebrows and wrinkled noses: The grammatical function of facial expression in relative clauses and related constructions. In *ASL in a Bilingual, Bicultural Context. Proceedings of the Second National Symposium on Sign Language Research and Teaching*, pages 65–74, Coronado, California, USA.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA.

[13] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[14] 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, Vienna, Austria.

[15] Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria.

[16] Emre Hakgüder. 2021. *Iconicity in Grammar: Typological Patterns in Sign Language Classifiers*. Ph.D. thesis, University of Chicago.

[17] Robert Hawkins, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663, Online. Association for Computational Linguistics.

[18] Annika Herrmann. 2013. *Modal and Focus Particles in Sign Languages: A Cross-linguistic Study*, second edition. De Gruyter Mouton.

[19] Al Amin Hosain, Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala, Jana Košecká, et al. 2020. Finehand: Learning hand shapes for American Sign Language recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 700–707, Buenos Aires, Argentina.

[20] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.

[21] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. Signbert: Pre-training

of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11087–11096, Montreal, Quebec, Canada.

[22] Jennifer Hu, Ethan Gotlieb Wilcox, Siyuan Song, Kyle Mahowald, and Roger P. Levy. 2026. What can string probability tell us about grammaticality? *Transactions of the Association for Computational Linguistics*, 14:124–146.

[23] Saki Imai, Lee Kezar, Laurel Aichler, Mert İnan, Erin Walker, Alicia Wooten, Lorna Quandt, and Malihe Alikhani. 2026. How pragmatics shape articulation: A computational case study in stem asl discourse. In *International Conference on Language Resources and Evaluation (LREC) 2026*, Palma, Mallorca, Spain.

[24] Mert İnan, 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.

[25] Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online.

[26] Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

[27] Serpil Karabüklü. 2024. Simultaneity of certainty in Turkish Sign Language (TİD). *Journal of Pragmatics*, 232:141–166.

[28] Serpil Karabüklü and Aslı Gürer. 2024. Prosody of focus in Turkish Sign Language. *Language and Cognition*, 16(4):1238–1271.

[29] Jonathan Keane and Diane Brentari. 2016. Fingerspelling: Beyond handshape sequences. In Marc Marschark and Patricia Spencer, editors, *The Oxford Handbook of Deaf Studies in Language: Research, Policy, and Practice*, pages 146–160. Oxford University Press.

[30] Lee Kezar, Nidhi Munikote, Zian Zeng, Zed Sehyr, Naomi Caselli, and Jesse Thomason. 2025. The American Sign Language knowledge graph: Infusing ASL models with linguistic knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7032–7044, Albuquerque, New Mexico.

[31] Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Berger, Jesse Thomason, Zed Sehyr, and Naomi Caselli. 2023. The Sem-Lex benchmark: Modeling ASL signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, 34, pages 1–10, New York, NY, USA.

[32] Lee Kezar, Zed Sehyr, and Jesse Thomason. 2025. Phonological representation learning for isolated signs improves out-of-vocabulary generalization. *Preprint*, arXiv:2509.04745. Version 1.

[33] Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 133–144, Amherst, MA. Association for Computational Linguistics.

[34] Cara Su-Yi Leong and Tal Linzen. 2024. Testing learning hypotheses using neural networks by manipulating learning data. *Preprint*, arXiv:2407.04593. Version 3.

[35] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*.

[36] Scott K Liddell. 1980. *American Sign Language Syntax*. Mouton de Gruyter.

[37] Scott K. Liddell. 1986. Head thrust in ASL conditional marking. *Sign Language Studies*, 52:244–262.

[38] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

[39] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1–4. Long Beach, CA.

[40] Xiaohan Ma, Rize Jin, and Tae-Sun Chung. 2024. Multi-channel spatio-temporal transformer for sign language production. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11699–11712, Torino, Italia.

[41] Evie Malaia, Joshua D Borneman, and Ronnie B Wilbur. 2018. Information transfer capacity of articulators in American Sign Language. *Language and Speech*, 61(1):97–112.

[42] Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium.

[43] Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models:

A case study on cross-dative generalization. *Preprint*, arXiv:2408.05086. Version 2.

[44] Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA.

[45] Carol Neidle, Judy Kegl, Dawn MacLaughlin, Benjamin Bahan, and Robert G. Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT press.

[46] Carol Neidle, Augustine Opoku, Carey Ballard, Konstantinos M. Dafnis, Evgenia Chroni, and Dimitri Metaxas. 2022. Resources for computer-based sign recognition from video, and the criticality of consistency of gloss labeling across multiple large ASL video corpora. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 165–172, Marseille, France. European Language Resources Association.

[47] Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022. ASL video corpora sign bank: Resources available through the american sign language linguistic research project (ASLLRP). *Preprint*, arXiv:2201.07899. Version 1.

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

[49] 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.

[50] Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

[51] Roland Pfau and Josep Quer. 2010. Nonmanuals: Their prosodic and grammatical roles. In Diane Brentari, editor, *Sign Languages*, pages 381–402. Cambridge University Press.

[52] Roland Pfau, Markus Steinbach, and Bencie Woll, editors. 2012. *Sign Language. An International Handbook*. De Gruyter Mouton, Berlin, Boston.

[53] Josep Quer, Roland Pfau, and Annika Herrmann. 2021. *The Routledge Handbook of Theoretical and Experimental Sign Language Research*. Routledge.

[54] Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, 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)*, page 8624–8641, Bangkok, Thailand.

[55] Wendy Sandler and Diane Lillo-Martin. 2006. *Sign Language and Linguistic Universals*. Cambridge University Press.

[56] Marcelo Sandoval-Castaneda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. Self-supervised video transformers for isolated sign language recognition. *Preprint*, arXiv:2309.02450. Version 1.

[57] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020. Adversarial training for multi-channel sign language production. In *Proceedings of the 31st British Machine Vision Virtual Conference (BMVC)*, Online, UK.

[58] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7881–7892, Online.

[59] Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

[60] Barbara Shaffer. 2004. Information ordering and speaker subjectivity: Modality in ASL. *Cognitive Linguistics*, 15:175–195.

[61] Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, Abu Dhabi, United Arab Emirates.

[62] Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

[63] Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4):1267–1301.

[64] Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA.

[65] Garrett Tanzer. 2025. FLEURS-ASL: Including American Sign Language in massively multilingual multitask evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6167–6191, Albuquerque, New Mexico.

[66] Garrett Tanzer and Biao Zhang. 2024. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. *arXiv preprint arXiv:2407.11144*.

[67] Sandrine Tornay, Necati Cihan Camgöz, Richard Bowden, and Mathew Magimai Doss. 2021. A phonology-based approach for isolated sign production assessment in sign language. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 102–106, Virtual Event, Utrecht, Netherlands.

[68] Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A large-scale, open-domain American Sign Language-English parallel corpus. In *Advances in Neural Information Processing Systems*, volume 36, pages 29029–29047.

[69] Clayton Valli and Ceil Lucas. 2000. *Linguistics of American Sign Language: An introduction*. Gallaudet University Press.

[70] Silvana C Veinberg and Ronnie B. Wilbur. 1990. A linguistic analysis of the negative headshake in American Sign Language. *Sign Language Studies*, 68:217–244.

[71] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

[72] Katharine L. Watson. 2010. WH-question in American Sign Language: Contributions of non-manual marking to structure and meaning. Master's thesis, Purdue University.

[73] Traci Patricia Weast. 2008. *Questions in American Sign Language: A Quantitative Analysis of Raised and Lowered Eyebrows*. Ph.D. thesis, The University of Texas at Arlington.

[74] Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[75] Ronnie B. Wilbur. 1994. Eyeblinks & ASL phrase structure. *Sign Language Studies*, 84(1):221–240.

[76] Ronnie B. Wilbur. 2011. Nonmanuals, semantic operators, domain marking, and the solution to two outstanding puzzles in ASL. *Sign Language & Linguistics*, 14(1):148–178.

[77] Ronnie B. Wilbur. 2021. Non-manual markers: Theoretical and experimental perspectives. In Josep Quer, Roland Pfau, and Annika Herrmann, editors, *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, pages 530–565. Routledge.

[78] Ronnie B. Wilbur and Cynthia Patschke. 1999. Syntactic correlates of brow raise in ASL. *Sign Language & Linguistics*, 2(1):3–41.

[79] Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

[80] Ryan Wong, Necati Cihan Camgöz, and Richard Bowden. 2025. SignRep: Enhancing self-supervised sign representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22804–22814, Honolulu, Hawaii, USA.

[81] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

[82] Tianyang Xu, Marcelo Sandoval-Castaneda, Karen Livescu, Greg Shakhnarovich, and Kanishka Misra. 2026. Cross-modal taxonomic generalization in (vision-) language models. *Preprint*, arXiv:2603.07474. Version 1.

[83] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

[84] Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. Both direct and indirect evidence contribute to dative alternation preferences in language models. In *Conference on Language Modeling*, Montreal, Canada.

[85] Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. 2021. Signed coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4950–4961, Online and Punta Cana, Dominican Republic.

[86] Kayo Yin, Chinmay Singh, Fyodor O Minakov, Vanessa Milan, Hal Daumé III, Cyril Zhang, Alex Xijie Lu, and Danielle Bragg. 2024. Asl stem wiki: Dataset and benchmark for interpreting stem articles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

[87] Biao Zhang, Garrett Tanzer, and Orhan Fırat. 2024. Scaling sign language translation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, pages 112018–114047, Red Hook, NY, USA.

[88] Inge Zwitserlood. 2012. Classifiers. In Roland Pfau, Markus Steinbach, and Bencie Woll, editors, *Sign Language. An International Handbook*, pages 158–186. De Gruyter Mouton, Berlin, Boston.

A. Complementary Results

Figure 4 shows the average surprisal differences (ΔSurprisal) between the matched and the mismatched translations for the training time cue-ablations (i.e., "controlled rearing").

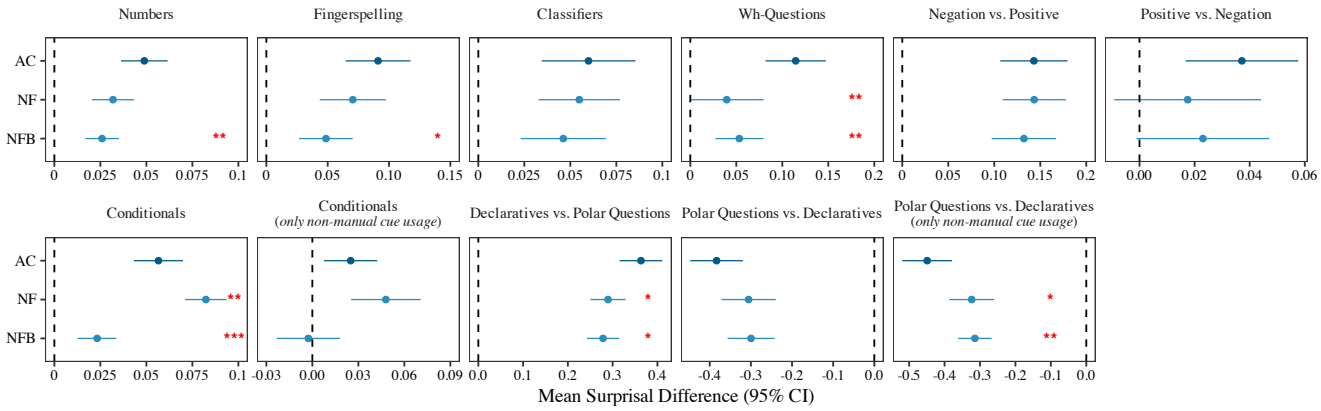

Figure 4. Average difference in surprisal of mismatched and matched sentences across phenomena and across **training ablations**. Error bars indicate 95% confidence intervals. Stars (*) indicate significance test results for comparing surprisal difference in a given cue ablation to surprisal difference in the all cues condition (AC). *: $p < .05$; **: $p < .01$; ***: $p < .001$