

Diagnosing Video Foundation Models for Single-Signer RGB-Only Auslan-Daily News Translation

Maoyang Li¹, Prashan Premaratne²[0000-0002-5750-3800] and Peter Vial³[0000-0001-9711-0115]

^{1,2,3} School of Electrical, Computer and Telecommunications Engineering,
University of Wollongong, North Wollongong, NSW, 2522, Australia
prashan@uow.edu.au

Abstract. Sign language translation (SLT) from video remains a challenging task because it requires models to connect fine-grained visual motion with spoken-language sentences. Recent progress in video foundation models suggests that stronger visual encoders may improve SLT, especially in low-resource settings. In this paper, we examine this assumption on the Auslan-Daily News split under a single-signer, RGB-only setting. We systematically compare three representative pipeline families, namely I3D, Image Swin-T, and VideoMAE, under a unified evaluation protocol. VideoMAE achieves the best performance but still remains clearly lower than the official benchmark. To better understand this gap, we further analyze model outputs, test sensitivity to temporal order and visual content, and compare several VideoMAE-based variants. Our results suggest that, in this controlled single-signer RGB-only Auslan-Daily News setting, the limitation cannot be explained by backbone choice alone. Instead, the main difficulty lies in how long video sequences are compressed and then carried into the translation stage under limited supervision. These findings indicate that stronger visual encoders alone may be insufficient for this setting, and that future work should pay closer attention to temporal modeling and visual-to-text bridging.

Keywords: Sign Language Translation, Video Foundation Models, Low-Resource Auslan Translation.

1 Introduction

Sign language is the primary means of communication for deaf and hard-of-hearing individuals, yet it is not widely understood by the hearing community [1]. Therefore, translating sign language videos into spoken language text becomes crucial. In realistic scenarios, the goal of sign language translation (SLT) is no longer limited to isolated word recognition or closed small scale vocabulary tasks but extends to sentence-level translation in open and content-rich domains. This substantially increases the difficulty of the task, especially on datasets like Auslan-Daily News split, where models not only have to handle natural sentence variations but also face real communication content [2, 3].

Remarkable progress in large-scale visual and video pre-training brings new opportunities for enhancing SLT. Models like Swin Transformer and VideoMAE have demonstrated strong transferability in general visual and video understanding tasks, and recent SLT work has begun to move towards larger-scale pre-training, stronger language models, and more open task scenarios [4-8]. Under this trend, it is reasonable to expect that better pre-trained video encoders should provide a stronger basis for SLT in low-resource settings.

However, our experiments on the Auslan-Daily News split do not fully support this expectation [2]. In the single-signer, RGB-only setting, stronger pre-trained backbones still underperform the official benchmark. This setting is already simpler than many real-world scenarios, there is no multi-person interaction in the video, and the background is clean. Therefore, this gap raises a more specific question, if strong pre-trained video models can provide useful visual representations, why are their benefits limited in low-resource SLT?

We treat this work as a controlled diagnostic study rather than a general evaluation of all Auslan SLT scenarios. To answer this question, we systematically compare three representative pipeline families on Auslan-Daily News split, namely I3D [9], Image Swin-T [4], and VideoMAE [5]. We evaluate their translation performance under a unified evaluation protocol, analyze their output length and vocabulary usage, and test their sensitivity to temporal order and visual content. We also examine several VideoMAE-based variants to study whether changing the bridge or decoder is enough to improve translation performance. The results show that the main difficulty is not backbone selection, but how temporal information is compressed and then carried into the translation stage. Based on these findings, we discuss what future Auslan SLT models may need in order to make better use of strong video representations.

2 Related Work

2.1 Sign Language Recognition and Translation

Sign Language Recognition. Sign language recognition (SLR) refers to identifying sign gestures or gloss sequences from video clips. With the development of deep learning, the performance of SLR has been significantly improved, and convolutional neural networks (CNNs) as well as Transformer-based architectures have become the main methods of the RGB-based SLR [10, 11]. In typical systems, CNN or 3D-CNN backbones are used to extract spatiotemporal visual features from sign video clips, while Transformer are employed to model long-range temporal dependencies in sign sequences [11, 12]. In datasets, besides widely studied benchmarks such as RWTH PHOENIX Weather and CSL, recent study has also explored recognition for Auslan, for example through the MM-WLAuslan dataset and related learning frameworks [13].

Sign Language Translation. Sign Language Translation (SLT) aims to convert

continuous sign language videos to spoken language sentences [14]. Early research typically used a gloss-based pipeline, in which intermediate gloss sequences were first predicted through SLR and then translated into the target spoken language [14]. Recently, some research has explored end-to-end or gloss-free translation methods, directly generating spoken-language text from video and reducing the reliance on manual gloss annotations [14]. In the architecture of models, most SLT systems adopt an encoder-decoder framework, where the visual encoder extracts spatiotemporal features from sign videos, then the text decoder generates corresponding sentences. At the same time, the availability of large-scale datasets has further promoted the development of continuous SLT research. For example, the How2Sign dataset has provided many sign language videos with corresponding spoken language transcripts, which provided important data resource for training translation models [15]. The recently proposed Auslan Daily dataset introduced Auslan SLT data collated from daily communication and news scenarios, making it possible to study Auslan SLT in more realistic communication environments [2]. However, translating continuous sign videos to spoken language remains challenging, especially in realistic and low-resource settings.

Sign Language Translation Challenges. Although research on SLT has made some progress in recent years, it still faces multiple challenges. First, SLT is inherently a long-sequences generation task, sign videos often contain many consecutive frames, and the generated sentences may involve long syntactic dependencies, making it difficult for sequence modeling [1]. Second, SLT requires accurate temporal and semantic alignment between visual signals and text semantics, that is, mapping continuous visual motions to corresponding words or phrases, a process that is often difficult to optimize [1, 16]. Third, many SLT datasets remain limited in scale and real-world scenarios often contain long-tail vocabulary, proper nouns, or misspelling, which further increase the difficulty of translation tasks [1, 2].

Visual Backbones for Video Understanding. Early research on video understanding mainly relied on convolutional architectures. I3D is a representative classic example that inflates 2D convolutional networks into 3D convolutional networks, directly learns spatiotemporal feature extractors from videos and establishes a strong baseline for action recognition [9]. Most recently, Transformer-based visual backbones gradually became an important direction in video modelling. Swin Transformer adopts hierarchical architecture and shifted windows mechanism, designed as a general visual backbone, and has strong performance in various visual tasks [4]. At the same time, self-supervised video pre-training has become an important route for learning transferable video representations. VideoMAE uses masked video modeling for video pre-training, emphasizes effective representation learning under limited data, and shows strong performance on multiple standard video benchmarks [5]. These backbones have demonstrated their effectiveness in action recognition, video

classification, and general video understanding tasks.

However, it is still unclear how well these backbones transfer to low-resource SLT. Unlike action recognition, SLT requires both visual encoding and sentence generation, which places additional demands on temporal compression and visual-text alignment. This gap motivates our study.

3 Experimental Setup

3.1 Task Setting and Dataset

In this study, we focus on RGB-only SLT on the Auslan-Daily News split under a single-signer setting. We use the official train/dev/test partition provided by Auslan-Daily and do not introduce any custom re-splitting or joint training with the Communication subset [2]. We choose the News subset as the main target of our study because it is more challenging in terms of vocabulary size, sentence complexity, and real-world variation, and it is also more closely aligned with our current benchmark goal [2]. All experiments use only RGB video as input. No pose, keypoint, or other additional modalities are used during either training or inference. This setting allows us to examine the effect of different visual backbones on translation performance in a controlled way.

3.2 Compared Pipelines and Visual Backbones

We compare three representative pipeline families. The first is an I3D feature pipeline, which extracts offline I3D features from videos and inputs the features into Transformer-based translation models as a classical video-feature baseline [9]. The second is an Image Swin-T pipeline. This method applies a frame-wise 2D Swin-T visual encoder to extract image features for each frame independently, followed by a temporal module and a text decoder for translation [4]. The third is a VideoMAE + time-rule pipeline, which is our main RGB SLT model. VideoMAE, as an end-to-end RGB visual backbone, combines explicit temporal compression to organize video features into a more suitable temporal representation for downstream translation [5]. In addition to these three main systems, we also examine VideoMAE-based variants such as RGBTQ and SpaMo. Since these methods mainly modify the temporal bridging module or the text-side architecture on VideoMAE backbone, rather than introducing a distinct backbone family, we do not include them in the main comparison table but discuss them in supplementary analyses and ablation experiments.

Figure 1 summarizes the three pipeline families compared in this study. I3D translates offline video features, Image Swin-T performs frame-wise 2D encoding followed by temporal modeling, and VideoMAE models video tokens directly before time-rule compression and decoding.

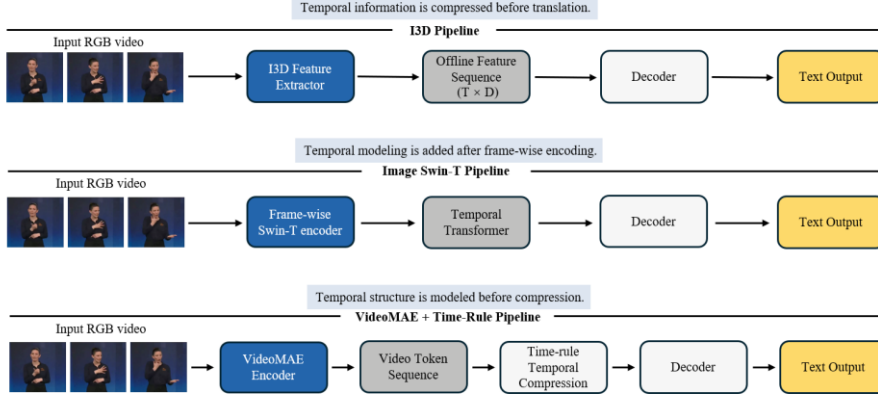


Fig. 1. Unified comparison of the three pipeline families

3.3 Training and Evaluation Protocol

To ensure fair comparison, we use a unified training and evaluation protocol in all main experiments. All inputs are uniformly sampled into 32 frames with a spatial resolution of 224×224 . Target text is uniformly encoded using SentencePiece tokenizer, with a vocabulary size of 5000. Before evaluation, both reference text and model predicted text are normalized using the UQ/Auslan-Daily style protocol, including lowercasing, removing punctuation, and whitespace normalization, to ensure consistency with the evaluation criteria and Auslan-Daily experiments. Translation quality is measured using BLEU-1/2/3/4 and ROUGE-L, where ROUGE-L calculates the F1 score of text matching based on the longest common subsequence [17, 18]. For model selection, we consistently choose the checkpoint with the best dev BLEU-4 score and report its corresponding performance on the test set.

4 Results and Analysis

The results in this section should be interpreted as a diagnostic analysis rather than as a claim of state-of-the-art performance. We use the gap between our strongest VideoMAE-based model and the official RGB benchmark as the starting point for examining why stronger pre-trained video encoders do not automatically lead to better low-resource SLT performance.

4.1 Backbone Comparison

Table 1 shows the results of different models with the official RGB benchmark on the Auslan-Daily News split under single-signer and RGB-only settings. In the official benchmark, single-signer RGB baselines reach test BLEU-4 scores between 2.26 and

2.68, where the highest BLEU-4 is obtained by SL-Luong (2.68) and the highest ROUGE-L is obtained by GASLT (22.01) [2]. In contrast, none of our three pipeline families surpasses the official benchmark.

In our systems, VideoMAE + time-rule is the strongest RGB-only pipeline which achieves the best test BLEU-4 score of 1.42. By comparison, I3D feature pipeline obtains a slightly higher BLEU-4 on the dev set than VideoMAE (1.76 vs 1.72), but its test BLEU-4 drops to 0.72. This suggests that offline video features can still provide useful spatiotemporal information, but they do not transfer reliably to sentence-level translation.

Table 1. Comparison with the official RGB benchmark on the Auslan-Daily News split under the single-signer RGB-only setting.

Category	Model / Pipeline	Visual backbone	Input type	Dev BLEU -4	Test BLEU -4	Test ROUGE -L	Notes
Prior work	SL-Luong	RGB baseline	RGB	–	2.68	16.14	Official benchmark
Prior work	SL-Transf	RGB baseline	RGB	–	2.52	14.93	Official benchmark
Prior work	TSPNet-Joint	RGB baseline	RGB	–	2.26	19.71	Official benchmark
Prior work	MMTLB	RGB baseline	RGB	–	2.31	18.90	Official benchmark
Prior work	GASLT	RGB baseline	RGB	–	2.56	22.01	Official benchmark
Ours	I3D feature pipeline	I3D-R50	Offline video features	1.76	0.72	10.44	Classical video baseline
Ours	Image Swin-T pipeline	Swin-T	Frame-wise RGB	0.68	0.44	12.02	2D image backbone + temporal modeling
Ours	VideoMAE + time-rule	VideoMAE-Base	End-to-end RGB	1.72	1.42	10.93	Main model / best pure RGB result

The results of Image Swin-T pipeline further show that frame-wise 2D visual encoding, followed by temporal modeling, is insufficient to support the current Auslan SLT task. The BLEU-4 scores on the dev set and test set are only 0.68 and 0.44, significantly lower than other two pipelines. Although Swin’s ROUGE-L is slightly higher than that of I3D and VideoMAE, the subsequent output analysis shows that this is mainly due to short and general high-frequency phrases rather than stable video-to-text mapping.

4.2 Output Length and Vocabulary Collapse

To better understand the BLEU-4 differences in Table 1, Table 2 reports the length and vocabulary statistics of the reference texts and model outputs on the test split. Sentence length alone does not explain translation quality. For BLEU-4, the key issue is whether a model can produce outputs whose local phrase patterns match the reference, that is, whether it can achieve sufficient higher-order n-gram overlap [17]. The statistics in Table 2 are therefore discussed in terms of length, vocabulary coverage, and output diversity.

VideoMAE is closest to the reference in sentence length, with an average length of 15.827 and a median of 15, compared with 16.217 and 15 for the reference text. I3D is slightly shorter overall, with an average length of 15.187 and a median of 12. Swin, by contrast, produces longer outputs, with an average length of 18.213 and a median of 18. This does not indicate better translation. Instead, it suggests that Swin tends to generate long generic sentences that do not match the reference content well. In BLEU-4, longer outputs can introduce more incorrect 4-grams when the generated content is poorly aligned with the reference, which helps explain why Swin obtains the lowest BLEU-4 despite producing longer sentences [17].

The vocabulary statistics show the same pattern. The reference text contains 2922 unique words, while I3D and VideoMAE reduce this to 903 and 1405, respectively. Both are well below the reference, but VideoMAE preserves a broader output vocabulary. Swin is much more restricted, producing only 43 unique words in total. In a news-domain task that includes named entities, long-tail words, and fixed expressions, such limited lexical coverage makes accurate translation difficult and lowers the chance of matching higher-order n-grams in the reference [2, 17]. This pattern matches Swin’s low BLEU-4 score of 0.44.

Output diversity is also closely related to the final score. Swin produces only 11 unique sentences over 700 test samples, and one sentence, “it is a lot of people who are not to be a lot of people in the country”, appears 526 times. This indicates that the model has collapsed to a very small set of repetitive templates and is no longer producing outputs that depend meaningfully on the input video. Under this condition, even relatively long outputs do not provide the sample-specific phrase structure needed for BLEU-4 matching. I3D and VideoMAE generate 576 and 492 unique sentences, respectively, and their most frequent predictions occur only 14 and 22 times. Both still show repetition, but the collapse is much less severe than in Swin, which is consistent with their higher BLEU-4 scores.

Comparing I3D and VideoMAE more closely, both produce sentence lengths that are broadly similar to the reference, but VideoMAE retains a larger vocabulary (1405 vs

903) and matches the reference median sentence length exactly. VideoMAE not only maintains a more suitable output length distribution, but also supports a wider range of lexical combinations, which increases the chance of matching local phrase patterns in the reference text. This is consistent with its higher test BLEU-4 (1.42 vs 0.72). I3D does not show the same degree of template collapse, but its smaller vocabulary and shorter median sentence length suggest stronger compression of the output semantics, which likely limits its ability to cover longer news sentences, named entities, and more complex expressions.

Table 2 shows that the three systems fail in different ways. Swin is dominated by severe vocabulary collapse and repetitive templates, which prevent it from producing effective sample-specific translations. I3D is limited more by reduced lexical coverage and semantic compression. And VideoMAE performs best because it reaches a better balance between output length, vocabulary coverage, and sentence diversity.

Table 2. Length and vocabulary statistics of reference text and output from different systems on the Auslan-Daily News test set.

Metric	Reference	I3D	Swin-T	VideoMAE
Avg sentence length	16.217	15.187	18.213	15.827
Median sentence length	15	12	18	15
Vocabulary size	2922	903	43	1405
Total tokens	11352	10631	12749	11079
Total sentences	700	700	700	700
Unique predicted sentences	—	576	11	492

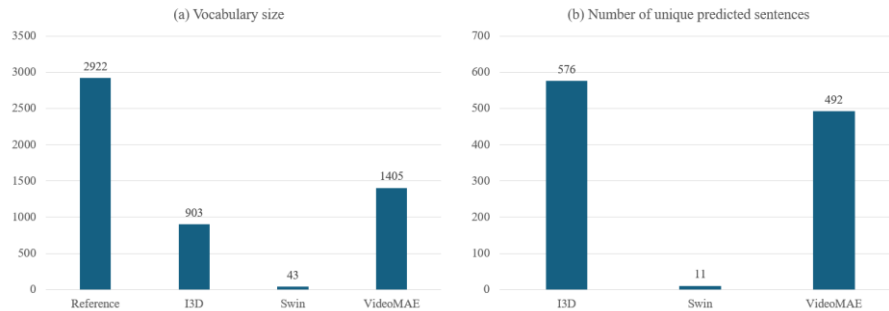


Fig. 2. Output-space collapse on the Auslan-Daily News test split. (a) Vocabulary size of the reference text and model outputs. (b) Number of unique predicted sentences generated by each model. All models produce substantially smaller output spaces than the reference text, with Swin showing the most severe collapse.

Figure 2 provides a more direct view of this output-space reduction. As shown in Figure 2(a), the output vocabularies of all three systems are substantially smaller than

the reference vocabulary, with Swin showing the most severe collapse. Figure 2(b) shows the same pattern at the sentence level. Swin produces only a very small number of distinct sentences, whereas I3D and VideoMAE remain more diverse but still fall short of the reference distribution.

Table 3 provides representative test cases to complement the aggregate statistics in Table 2 and Fig. 2. To keep the table readable, we report the key reference cues rather than the full normalized sentences, and summarize the main behaviour of each model. These cases show that VideoMAE can sometimes preserve the full sentence or retain topic-level cues, but it often loses named entities and event-specific details in longer news sentences. Swin-T shows a more severe template-collapse pattern, repeatedly producing the same generic sentence across unrelated inputs, while I3D generates more diverse but often semantically drifting news-style outputs.

Table 3. Representative translation cases from the Auslan-Daily News test split.

Case	Reference cue	Case-level observation
1	the news in auslan will be back next week	VideoMAE exactly reproduces the reference, while I3D and Swin-T generate unrelated news-style sentences. Successful VideoMAE case.
2	Queen’s funeral; Westminster Abbey; public preparations	VideoMAE keeps the funeral topic, but misses detailed event information. Swin-T produces a repeated generic template. Topic preserved, event details lost.
3	Harry Souttar; Mat Ryan; World Cup clean sheet	VideoMAE captures the football topic, but misses named entities and specific facts. Topic preserved, named entities lost.
4	US embassy in Sudan; evacuation; Khartoum fighting	VideoMAE generates fluent political text but changes the event semantics. Swin-T again collapses to the repeated template. Fluent but semantically wrong.

4.3 Visual Sensitivity Analysis

To examine how strongly each system depends on the video input, we evaluate all models under three conditions. In the normal setting, the original input sequence is used. In the shuffled setting, the frames of each sample are randomly permuted while the visual content is kept unchanged. In the zero setting, the input is replaced with an all-zero tensor of the same shape. The results are reported in Table 4. This analysis is meant to separate sensitivity to visual content from sensitivity to temporal order, and to identify cases where a model has largely reduced to a text-side template generator.

The I3D pipeline is highly sensitive to whether visual content is present but shows almost no sensitivity to temporal order. Its BLEU-4 stays at 0.72 under both the normal and shuffled conditions, and ROUGE-L is also nearly unchanged. Once the input is zeroed out, BLEU-4 drops to 0.02 and ROUGE-L to 1.66. The average output length shrinks from about 15 words to 6, the output vocabulary drops from about 900 words to 4, and all test samples collapse to a single sentence. This indicates that I3D does rely on video content, but it uses overall visual content rather than fine-grained temporal

structure. It can distinguish whether usable visual features exist, but it does not appear to benefit much from preserving the original temporal order.

The Swin pipeline shows a different failure mode. Its BLEU-4 remains between 0.44 and 0.47 across the normal, shuffled, and zero conditions, and ROUGE-L stays around 12 in all three cases. Sentence length, vocabulary size, and output diversity also change very little. Even when the video input is removed entirely, the model produces almost the same outputs in the normal setting. This indicates that the Swin pipeline relies weakly on the input video and has largely collapsed into a small set of generic sentence templates.

VideoMAE behaves differently from I3D and Swin. Under normal settings, it achieves the strongest BLEU-4 among the three pipelines, at 1.42. When the temporal order is destroyed, BLEU-4 drops to 0.35, and under zero input it further drops to 0.29. Since BLEU-4 depends on higher-order n-gram overlap with the reference text [17], this drop indicates that the loss of temporal structure weakens local phrase-level matching. At the same time, the perturbation results also reveal a limitation. After shuffling, the average output length rises sharply to 25.429 words and the vocabulary size remains large, but the score drops substantially. The model still produces fluent sentences, but these sentences are no longer aligned with the input content. Under zero input, the vocabulary collapses to 12 words and all samples reduce to a single output sentence. VideoMAE therefore makes more meaningful use of video content than the other two systems, but the learned representation is still not robust enough to support stable translation in the current low-resource Auslan SLT setting.

Table 4 points to three different modes. I3D depends on the presence of visual input but is largely insensitive to frame order. Swin remained almost unchanged under all three input conditions, which indicates severe template collapse, with only weak dependence on the input video. VideoMAE is sensitive to both visual content and temporal order, but its performance still degrades once either is disrupted.

Table 4. Performance under different input perturbation conditions on the Auslan-Daily News test split.

Model	Input condition	B4	ROUGE-L	Avg length	Vocab size	Unique sentences
I3D	Normal	0.72	10.44	15.187	903	576
I3D	Shuffled	0.72	10.42	15.237	906	576
I3D	Zero	0.02	1.66	6.000	4	1
Swin	Normal	0.44	12.02	18.213	43	11
Swin	Shuffled	0.44	12.02	18.214	43	11
Swin	Zero	0.47	12.59	18.000	14	1
VideoMAE	Normal	1.42	10.93	15.827	1405	492
VideoMAE	Shuffled	0.35	9.61	25.429	1520	535
VideoMAE	Zero	0.29	10.44	14.000	12	1

Table 5 reports several derived metrics based on the output statistics in Table 2 and the perturbation results in Table 4. Vocabulary retention measures the ratio between the model output vocabulary and the reference vocabulary. Sentence diversity measures the

proportion of unique predictions among the 700 test samples. Template dominance measures how often the most frequent prediction appears in the test set. Temporal sensitivity is computed as the BLEU-4 difference between the normal and shuffled settings, while visual dependence is computed as the BLEU-4 difference between the normal and zero-input settings.

Table 5. Quantitative diagnostic metrics for output collapse and input sensitivity.

Metric	I3D	Swin-T	VideoMAE
Vocabulary retention (%)	30.9	1.5	48.1
Sentence diversity (%)	82.3	1.6	70.3
Template dominance (%)	2.0	75.1	3.1
Temporal sensitivity (Δ BLEU-4)	0.00	0.00	1.07
Visual dependence (Δ BLEU-4)	0.70	-0.03	1.13

These derived metrics confirm the qualitative patterns above: Swin-T shows severe template collapse and weak visual dependence, I3D relies on visual content but not temporal order, and VideoMAE shows the strongest temporal and visual dependence but still does not reach benchmark-level translation quality.

Figure 3 gives a more direct view of how the three pipelines respond to input perturbations. In Figure 3(a), I3D remains nearly unchanged between the normal and shuffled conditions but collapses under zero input, while Swin shows almost no difference across all three conditions. VideoMAE drops once temporal order is destroyed. Figure 3(b) shows a similar trend in ROUGE-L, although the change is less pronounced than in BLEU-4. This suggests that temporal perturbation affects higher-order translation matching more strongly than broad lexical overlap.

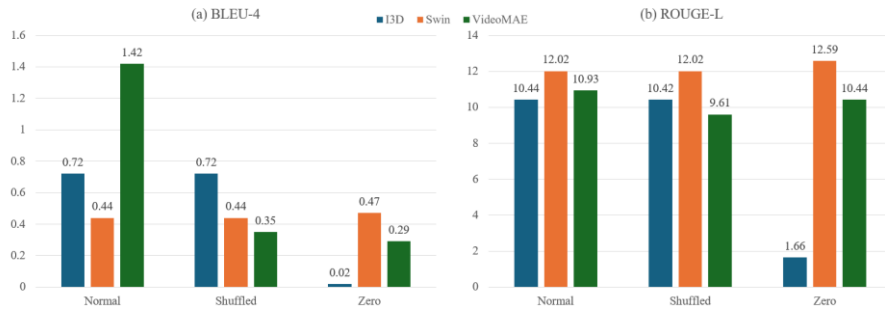


Fig. 3. Performance under input perturbations on the Auslan-Daily News test split. (a) BLEU-4 under normal, shuffled, and zero input conditions. (b) ROUGE-L under the same three conditions. I3D is sensitive to the presence of visual content but not to frame order, Swin changes little across all conditions, and VideoMAE is more sensitive to temporal disruption.

4.4 Additional Evidence on Transfer Limitations

Table 6 compares several variants built on VideoMAE backbone. Unlike Table 1, these models do not differ in the visual encoder itself. Instead, they differ in how the visual features are compressed and passed to the decoder. This comparison therefore helps clarify whether the current limitation mainly comes from the backbone or from the translation interface following backbone.

Table 6. Comparison of VideoMAE-based variants on the Auslan-Daily News split.

Variant	Visual backbone	Bridge / Decoder	Dev B4	Test B4	Test ROUGE-L	Notes
VideoMAE + time-rule	VideoMAE-Base	explicit time-rule compression + Transformer decoder	1.7201	1.42	10.93	Main model
VideoMAE + RGBTQ	VideoMAE-Base	learned temporal query bottleneck	1.3676	0.85	9.91	Replaces fixed compression
VideoMAE + SpaMo v1	VideoMAE-Base	spatial-motion aggregation bottleneck	1.6847	1.29	11.13	Stronger bridge, but still below benchmark
SpaMo v4	VideoMAE-Base	larger bridge/decoder	0.0769	–	–	Very weak
SpaMo v5	VideoMAE-Base	larger bridge/decoder	0.5594	–	–	Collapsed later
SpaMo v6	VideoMAE-Base	DAPT + larger bridge/decoder	0.1802	–	–	Collapsed later

The results show that changing the bridge or decoder does not automatically improve translation quality. Among the three main variants, VideoMAE + time-rule remains the strongest, reaching BLEU-4 scores of 1.7201 on the dev set and 1.42 on the test set. RGBTQ drops to 1.3676 on dev and 0.85 on test, while SpaMo v1 reaches 1.6847 on dev but only 1.29 on test. In other words, even with the same pre-trained VideoMAE backbone, different downstream designs lead to different results, and none of them closes the gap to the official benchmark.

The training logs lead to the same conclusion. More complex bridge and text-side designs do not appear to be more reliable in this low-resource setting. SpaMo v5 reaches a best dev BLEU-4 of only 0.5594 before collapsing to 0, and SpaMo v6 peaks at 0.1802 before also dropping to 0. Even SpaMo v4 remains far below a usable range. These results suggest that the current limitation is not resolved simply by changing the bridge or decoder on top of the same VideoMAE backbone. The downstream translation interface remains a major source of performance variation.

5 Discussion

5.1 Temporal Compression and Fine-Grained Motion

The three main pipelines differ in how they handle temporal information, but all of them reduce it at some point in the model. In the I3D pipeline, temporal patterns are modeled first by 3D convolutions, and the full video is then converted into a fixed sequence of offline features [9]. The translation model therefore works with a compressed $T \times D$ representation instead of the original RGB frames. In the Image Swin-T pipeline, Swin itself is a frame-wise 2D encoder, so motion across frames is not modeled inside the visual backbone [4]. Temporal modeling is added only after per-frame features have already been extracted. VideoMAE + time-rule is stronger at the visual stage because VideoMAE models temporal structure directly over video tokens [5]. Even so, the sequence is later reduced by the time-rule module before decoding. As a result, the question is not only whether the model can capture motion, but also whether useful temporal detail survives the later compression stage.

The risk of losing fine temporal detail is therefore different across the three pipelines. Swin is the most fragile because it begins with independent frame encoding. Short motion transitions and local movement patterns must be recovered later from features that were not designed to preserve motion in the first place. I3D does model space and time jointly, but the downstream translation model only receives offline features, not the original video. Once fine detail has been compressed during feature extraction, it cannot be recovered later. VideoMAE + time-rule preserves temporal structure better during encoding, but it still shortens the sequence before translation. This creates a different risk, distinct local motions may be merged into the same token once the sequence is compressed.

It is important because sentence-level SLT depends on visual details that are easy to lose under temporal reduction [1]. In the current Auslan-Daily News setting, the most vulnerable information includes handshape variation, short motion transitions, fingerspelling, and local continuous movement. These cues depend on short temporal windows rather than on single-frame appearance or coarse action categories. Frame-wise 2D encoding weakens short motion cues. Offline fixed features make it harder for the translation model to retain fine hand detail. Strong temporal compression can further merge several distinguishable local motions into one representation. The key point is that all three pipelines reduce temporal detail at different stages, while sentence-level translation depends on exactly those local details.

5.2 Low-Resource and Long-Tail Difficulty

The difficulty of Auslan-Daily News does not come only from limited training data. The more important issue is the mismatch between the amount of supervision and the complexity of the task. For the current RGB end-to-end models, the News split does not provide enough supervision for an open, sentence-level translation problem with a strong long-tail distribution [2]. This means that the model must learn video-to-text mapping from limited data while also handling a large output space and substantial

linguistic variation. Under such conditions, general visual backbones can still learn basic video representations, but they are much harder to adapt into stable representations for translation.

This problem becomes more severe in the News domain. Compared with daily communication, news sentences are typically longer, denser in information, and more likely to contain low-frequency but important named entities and event-specific expressions [1, 2]. These elements often carry the main meaning of the sentence, but they are not seen often enough during training. Once the model cannot learn them reliably, it is more likely to fall back on common phrases and familiar sentence patterns to maintain fluent output, rather than generate translations that match the input content. This also helps explain why, although the systems in Section 4 fail in different ways, they all show some degree of reliance on frequent expressions.

Models such as VideoMAE and Swin are pre-trained to learn general visual or video representations [4, 5]. Such representations help the model process the input video, but they do not automatically become suitable intermediate semantic representations for translation. That step still requires enough downstream supervision to learn how visual features should align with words, phrases, and sentence-level meaning. At the current data scale, this adaptation is still insufficient.

Seen in this way, the impact of low-resource training is not only on the final performance ceiling but also changes the kind of mapping the model can learn. When supervision is limited, the model is more likely to learn frequent sentence patterns and coarse semantic regularities than stable mappings for low-frequency content and sentence-level meaning.

5.3 Visual-Language Misalignment and Translation Bottlenecks

The difference between SLT and action recognition is not only in the output format, but in what the model ultimately needs to learn. Action recognition only needs a class label, so the video can be compressed into a coarse global representation [9]. SLT is different. It has to turn a continuous video into a full sentence. For this task, it is not enough to keep only a rough sense of what the video is about. The model needs local temporal information and semantic boundaries that can support word-by-word generation [1, 14].

The current problem is not that the backbone fails to understand the video, but that its representation is still not stable enough for the translation stage. It does not become a stable representation that the decoder can use effectively. Section 4 already shows that VideoMAE is clearly stronger than Swin, and that I3D can also generate partly relevant short sentences. This means the models do learn some visual content, but not enough for stable translation. The real difficulty is that raw video representations are long and dense, while the decoder needs shorter and clearer units that can be mapped to words and phrases. When this step fails, the model tends to keep only coarse meaning or fall back on frequent language templates.

More specifically, the difficulty lies in the temporal bottleneck and the visual-language bridge. Whether the model uses offline I3D features or later modules such as time-rule, RGBTQ, and SpaMo on top of VideoMAE, the underlying problem is the same, how to compress a long video into a short but useful translation representation.

The current results suggest that this step is still weak. Even VideoMAE + time-rule, which performs best in our study, remains clearly below the benchmark. The issue is not only backbone strength, but whether enough translation-relevant information remains after compression.

Once this step is unstable, the decoder becomes more likely to rely on language priors instead of visual evidence. This tendency is already visible in the Swin results in Section 4, and the SpaMo variants. The model can learn fluent high-frequency sentence patterns, but these outputs do not necessarily match the input video. The harder step is turning long visual sequences into shorter units that still preserve the information needed for sentence generation.

5.4 Implications for Future Auslan Translation Models

Future Auslan translation models need improvements in three areas. First, temporal compression should be more adaptive. The current systems can process motion, but they lose too much temporal detail before translation begins. Instead of reducing every video to the same compressed temporal representation, future models may need to retain more intermediate information for longer or motion-dense clips. This is important because sentence-level SLT requires visual representations that can support translation rather than only coarse video understanding [1]. Second, the visual-to-text bridge needs to be more stable. Our results suggest that simply replacing the decoder or increasing its capacity is not enough. The compressed visual representation must remain useful for text generation after temporal reduction. Future models should therefore focus on producing intermediate representations that are shorter than the original video sequence but still preserve translation-relevant semantic cues. This direction is consistent with the encoder-decoder formulation of SLT, where visual features have to support spoken-language sentence generation [14]. Third, training objectives should be designed more directly for low-resource SLT. General video pre-training can provide useful visual representations, and VideoMAE is an example of a data-efficient self-supervised video pre-training framework [5]. However, our results suggest that such pre-training does not automatically produce representations that are well suited to sentence-level translation. Future baselines should therefore encourage the temporal bridge and decoder to depend more strongly on visual evidence, rather than allowing the decoder to fall back on frequent sentence patterns. These directions point to the same practical need: future models should preserve more temporal information, build a more reliable visual-to-text interface, and train the translation stage closer to the actual SLT objective.

5.5 Limitations

This study has several limitations. First, all experiments are conducted on the Auslan-Daily News split under the single-signer RGB-only setting. The results therefore do not directly establish whether the same failure modes appear on other sign language datasets, such as PHOENIX-2014T or How2Sign, or in multi-signer scenes. Second, our perturbation tests focus on frame shuffling and zero input. They reveal whether models depend on temporal order and visual content, but they do not provide direct

token-level alignment between signs and generated words. Third, the VideoMAE-based variants studied here modify the bridge or decoder, but they do not introduce a fully new translation architecture. The contribution of this paper is therefore diagnostic rather than model-driven.

6 Conclusion

This paper studies visual backbone transfer in low-resource Auslan SLT under the single-signer RGB-only Auslan-Daily News setting. VideoMAE performs best among the compared pipelines but still remains below the official RGB benchmark. This gap indicates that stronger video encoders alone are not sufficient in this controlled setting. Our analyses suggest that the main limitation lies in the translation interface after visual encoding, especially how temporal information is compressed and transferred to the decoder. These findings highlight that sentence-level Auslan SLT requires not only strong visual representations, but also a reliable temporal and visual-to-text interface.

References

- [1] Z. Liang, H. Li, J. Chai, Sign language translation: A survey of approaches and techniques, *Electronics*, 12 (2023) 2678.
- [2] X. Shen, S. Yuan, H. Sheng, H. Du, X. Yu, Auslan-daily: Australian sign language translation for daily communication and news, *Advances in Neural Information Processing Systems*, 36 (2023) 80455-80469.
- [3] B. Zhang, G. Tanzer, O. Firat, Scaling sign language translation, *Advances in Neural Information Processing Systems*, 37 (2024) 114018-114047.
- [4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012-10022 (2021).
- [5] Z. Tong, Y. Song, J. Wang, L. Wang, Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, *Advances in Neural Information Processing Systems*, 35 (2022) 10078-10093.
- [6] G. Tanzer, M. Shengelia, K. Harrenstien, D.C. Uthus, Reconsidering sentence-level sign language translation, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6262-6287 (2024).
- [7] J. Gong, L.G. Foo, Y. He, H. Rahmani, J. Liu, Llms are good sign language translators, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18362-18372 (2024).
- [8] J. Ye, X. Wang, W. Jiao, J. Liang, H. Xiong, Improving gloss-free sign language translation by reducing representation density, *Advances in Neural Information Processing Systems*, 37 (2024) 107379-107402.
- [9] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299-6308 (2017).
- [10] H. Hu, W. Zhao, W. Zhou, Y. Wang, H. Li, SignBERT: Pre-training of hand-model-aware

representation for sign language recognition, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11087-11096 (2021).

[11] Y. Zhang, X. Jiang, Recent Advances on Deep Learning for Sign Language Recognition, Computer Modeling in Engineering & Sciences, 139 (2024) 2399--2450.

[12] A. Ghadami, A. Taheri, A. Meghdari, A transformer-based multi-stream approach for isolated iranian sign language recognition, arXiv preprint arXiv:2407.09544, (2024).

[13] X. Shen, H. Du, H. Sheng, S. Wang, H. Chen, H. Chen, Z. Wu, X. Du, J. Ying, R. Lu, MM-WLAuslan: multi-view multi-modal word-level Australian sign language recognition dataset, Advances in Neural Information Processing Systems, 37 (2024) 69700-69715.

[14] N.C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10023-10033 (2020).

[15] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, X. Giro-i-Nieto, How2sign: a large-scale multimodal dataset for continuous american sign language, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2735-2744 (2021).

[16] Y. Said, S. Boubaker, S.M. Altowaijri, A.A. Alsheikhy, M. Atri, Adaptive transformer-based deep learning framework for continuous sign language recognition and translation, Mathematics, 13 (2025) 909.

[17] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318 (2002).

[18] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, Text summarization branches out, pp. 74-81 (2004).