Think in Latent Thoughts: A New Paradigm for Gloss-Free Sign Language Translation

Yiyang Jiang[1] **Li Zhang**[1] **Xiaoyong Wei**[1,2*] **Qing Li**[1]
[1]The Hong Kong Polytechnic University [2]Sichuan University
yiyang.jiang@connect.polyu.hk
{zanly20.zhang, cs007.wei, qing-prof.li}@polyu.edu.hk

Abstract

Many Sign language translation (SLT) systems quietly assume that brief chunks of signing map directly to spoken-language words. That assumption breaks down because signers often create meaning on the fly using context, space, and movement. We revisit SLT and argue that it is mainly a cross-modal reasoning task, not just a straightforward video-to-text conversion. We thus introduce a reasoning-driven SLT framework that uses an ordered sequence of latent thoughts as an explicit middle layer between the video and the generated text. These latent thoughts gradually extract and organize meaning over time. On top of this, we use a plan-then-ground decoding method: the model first decides what it wants to say, and then looks back at the video to find the evidence. This separation improves coherence and faithfulness. We also built and released a new large-scale gloss-free SLT dataset with stronger context dependencies and more realistic meanings. Experiments across several benchmarks show consistent gains over existing gloss-free methods. Our code and data are available at https://github.com/fletcherjiang/SignThought.

1 Introduction

Sign language translation (SLT) is critically important both as a vital assistive technology for connecting Deaf and hard-of-hearing communities and as a challenging multimodal task within natural language processing (Bragg et al., 2019). Research in this area has evolved significantly, progressing from early methods that approached SLT as a gloss-level classification problem over video segments (Koller et al., 2015) to more recent formulations as a gloss-free, video-to-text sequence transduction task (Lin et al., 2023; Chen et al., 2024a), tackled by specialized architectures or multimodal large language models (Wong et al., 2024; Chen et al., 2024b; Gong et al., 2024). While these efforts have yielded encouraging outcomes, substantial room for improvement remains.

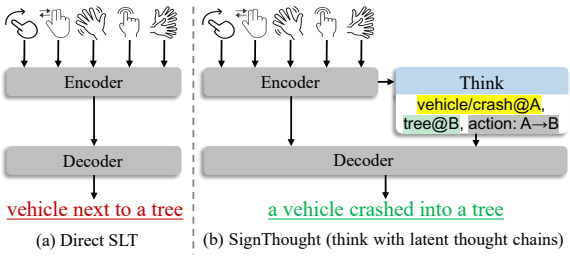


Figure 1: Latent thoughts clarified the context to disambiguate signs and guide decoding toward the correct object interaction in gloss-free SLT.

A key limitation lies in a long-standing misconception in model design: treating sign video segments as if they were directly mappable to static spoken-language words or phrases (Yin et al., 2021; Camgoz et al., 2018; Bragg et al., 2019). This view originates from an early emphasis on the *Frozen Lexicon*, i.e., the finite inventory of conventionalized signs listed in glossaries, while largely overlooking *Productive Forms* (Forster et al., 2014). Productive forms are constructed on-the-fly through classifiers, spatial grammar, and motion modulation, forming an open-ended and context-governed system that cannot be enumerated as discrete lexical entries (Zwitserlood, 2012). For instance, a single "vehicle" handshape may express meanings such as "park," "crash," or "drive" purely through variations in movement and spatial configuration (Zwitserlood, 2012; Ngo et al., 2008). As a consequence, meaning in sign language is often not explicitly encoded in fixed symbolic units but dynamically *generated* from motion, space, and discourse context (Zwitserlood, 2012; Cormier et al., 2012; Xiao-Yong Wei, 2013). SLT is therefore a generative, context-dependent reasoning problem, not just segment alignment or symbol substitution (Yin et al., 2021; Bragg et al., 2019; Gong et al., 2024; Jiang et al., 2024).

* Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9993–10012

From this view, SLT naturally aligns with recent advances in text reasoning techniques, where models explicitly maintain intermediate semantic states to support multi-step meaning composition under weak or implicit supervision (Wei et al., 2022; Yao et al., 2022). However, unlike text-only reasoning, SLT reasoning must operate across heterogeneous visual and linguistic modalities (Mitra et al., 2024). The absence of discrete reasoning primitives and explicit intermediate representations makes existing reasoning techniques difficult to transfer directly, leaving multimodal reasoning in SLT largely unexplored (Yin et al., 2021). We therefore position this work as a pilot study toward understanding how explicit reasoning can be modeled in gloss-free SLT, with the goal of identifying fundamental challenges and exploring principled solutions (Bragg et al., 2019; Jiang et al., 2026b). This paradigm offers three key advantages that tackle the reasoning challenges as:

- **Latent Thought Abstraction.** We introduce latent thoughts as an explicit intermediate semantic interface between sign videos and text generation. Rather than compressing all semantic information into opaque encoder features, the model maintains an ordered set of latent reasoning states that progressively distill and organize meaning from long, continuous visual streams, thereby bridging the mismatch between continuous visual evidence and discrete reasoning primitives.

- **Plan–Ground Decoupling.** Reasoning and grounding are explicitly separated through a plan-then-ground generation mechanism: the model first determines *what* semantic content to express by reasoning over latent thoughts, and only then decides *where* to retrieve the corresponding visual evidence. This decoupling alleviates the strong entanglement between semantic decision-making and evidence retrieval in existing SLT systems, improving controllability and grounding faithfulness.

- **Traceable Evidence Alignment.** Latent thoughts function not only as internal planning states but also as traceable anchors that align generated text with specific temporal regions of the input video, enabling explicit evidence attribution and more faithful translations, while a newly constructed large-scale gloss-free SLT dataset provides the essential empirical basis for studying and validating such reasoning behaviors.

2 Related Work

Sign Language Recognition and Translation. Sign Language Recognition (SLR) maps sign videos to gloss sequences, evolving from isolated (ISLR) (Albanie et al., 2020; Li et al., 2020a) to continuous settings (CSLR), with typical pipelines extracting visual features from RGB or pose-based inputs (Chen et al., 2022b), modeling temporal dynamics via RNN/LSTM or Transformer architectures (Camgoz et al., 2018, 2020), and decoding glosses using HMMs (Koller et al., 2017) or CTC objectives (Cheng et al., 2020). While effective, such systems are computationally intensive and limited in exploiting higher-level linguistic context. Sign Language Translation (SLT) further translates sign videos into spoken or written language, where most methods remain gloss-based by cascading SLR with translation or jointly optimizing both tasks (Camgoz et al., 2020; Zhou et al., 2021). In contrast, gloss-free SLT directly learns video-to-text mappings using Transformer-based or variational models (Li et al., 2020b; Camgoz et al., 2018; Tu and Weng, 2026); although recent advances leveraging large-scale multimodal pretraining (Li et al., 2025) and large language model adaptation (Wong et al., 2024) improve translation fluency, they introduce substantial computational overhead and increased dependence on external corpora.

Reasoning and Latent Thoughts. Chain-of-Thought (CoT) prompting enhances multi-step reasoning in large language models (LLMs) by eliciting intermediate reasoning steps (Wei et al., 2023; Zhou et al., 2024b, 2025), with extensions such as self-consistency (Wang et al., 2023), planning, search, and tool-augmented reasoning further improving robustness and capability (Yao et al., 2023; Xie et al., 2025; Zeng et al., 2025). Beyond general reasoning tasks, LLMs have also been applied to personalized recommendation assistants with memory and reasoning (Huang et al., 2026, 2025). CoT has also been explored in multimodal grounding and video reasoning (Zhang et al., 2024). More recently, reasoning has shifted from discrete language tokens to continuous latent spaces, where hidden states are reused as "latent thoughts" through feedback or iterative computation (Hao et al., 2025), with formal analyses and surveys

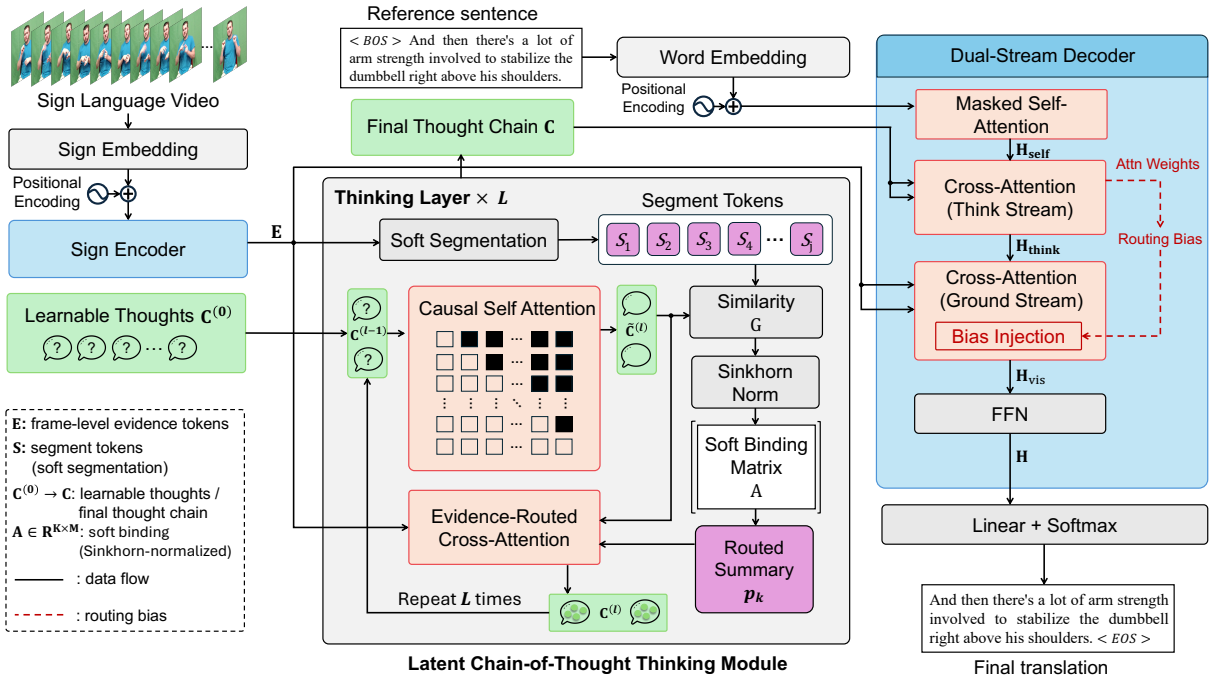


Figure 2: An overview of the SignThought framework, which consists of three parts: (i) a sign encoder that maps the input video $\mathbf{X}$ into dense evidence features $\mathbf{E}$; (ii) a Latent Chain-of-Thought module that iteratively updates K learnable thought slots and performs structured routing (segmentation $\rightarrow$ Sinkhorn-style binding $\rightarrow$ routed retrieval, with an optional monotonic bias) to distill $\mathbf{E}$ into an ordered thought chain $\mathbf{C}$; and (iii) a plan-then-ground dual-stream decoder that first attends to $\mathbf{C}$ for semantic planning and then grounds each token on $\mathbf{E}$ under thought-guided temporal priors to generate the spoken-language translation end-to-end.

clarifying the relationship between CoT and latent-thought paradigms (Xu and Sato, 2025). Notably, existing latent-thought approaches primarily operate on token-level embeddings within unimodal LLMs, whereas our Latent Thoughts are video-conditioned latent slots updated via attention over sign-video features, explicitly serving as a cross-modal interface bridging visual evidence and decoding in gloss-free SLT.

3 Method

We study gloss-free sign language translation (SLT), where the supervision consists solely of sentence-level translations without any intermediate gloss annotations. Given a sign video clip, we aim to generate a target spoken-language sentence. Let a sign clip be represented as a sequence of visual observations $\mathbf{X} = \{x_t\}_{t=1}^{T_s}$ (pre-extracted features). A binary mask $\mathbf{m} \in \{0,1\}^{T_s}$ indicates valid time steps when clips are padded. The target translation is a token sequence $\mathbf{Y} = \{y_t\}_{t=1}^{T_t}$ with $y_t \in \mathcal{V}$, where $\mathcal{V}$ is the text vocabulary. We assume a dataset of paired samples $\mathcal{D} = \{(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})\}_{n=1}^{N}$.

Crucially, no gloss sequence or sign-token alignment is provided. We learn a conditional sequence model that maximizes the likelihood of the target sentence given the sign clip:

$$p_\theta(\mathbf{Y} \mid \mathbf{X}) = \prod_{t=1}^{T_t} p_\theta(y_t \mid y_{<t}, \mathbf{X}). \quad (1)$$

In addition to the standard autoregressive formulation, our method introduces a set of K ordered latent thought states $\mathbf{C} = \{c_k\}_{k=1}^{K}$ that serve as an intermediate semantic plan distilled from $\mathbf{X}$ and consumed by the decoder during generation.

3.1 Sign Encoder

Given a sign video clip, we represent it as a feature sequence $\mathbf{X} = \{x_t\}_{t=1}^{T_s}$, where each $x_t \in \mathbb{R}^{d_x}$ is a pre-extracted visual feature vector from a pre-trained Inception network at time step t, and T_s is the padded sequence length. Following the feature setting of (Voskou et al., 2021), we initialize the visual feature extractor from the open-sourced Inception network and remove gloss-dependent supervision during feature learning. The extractor is trained with a sentence-level contrastive objective using only paired sign video-sentence data. We use a binary mask $\mathbf{m} \in \{0,1\}^{T_s}$ to indicate valid (non-padded) positions. We first project the raw

per-frame features into the model space via a learnable embedding layer, i.e., $\mathbf{E}^{(0)} = \mathrm{Embed}(\mathbf{X})$, where $\mathrm{Embed}(\cdot)$ is implemented as a linear projection followed by normalization and dropout. This step unifies sign features into a shared representation space for subsequent sequence modeling.

To inject order information, we add positional encoding to obtain $\hat{\mathbf{E}}^{(0)}$, with dropout applied afterward. We then stack N_{enc} encoder layers to model long-range dependencies and local motion patterns:

$$\hat{\mathbf{E}}^{(n)} = \mathrm{EncBlock}^{(n)}(\hat{\mathbf{E}}^{(n-1)}, \mathbf{m}), \ n = 1, \ldots, N, \tag{2}$$

where each $\mathrm{EncBlock}$ follows a Conformer-style design, combining multi-head self-attention (masked by $\mathbf{m}$) with a lightweight convolutional module to capture local temporal dynamics, together with feed-forward sublayers and residual normalization. The final encoder output is $\mathbf{E} = \hat{\mathbf{E}}^{(N)} = \{e_t\}_{t=1}^{T_s}$. The resulting sequence $\mathbf{E}$ serves as the frame-level evidence source for both the Latent CoT thinking module (Sec. 3.2) and the dual-stream decoder (Sec. 3.3).

3.2 Thinking in Latent Chain-of-Thought

The objective of the thinking module is to transform the dense encoder representations $\mathbf{E}$ into a compact and ordered set of K latent thought states $\mathbf{C} = \{c_k\}_{k=1}^{K}$, which collectively serve as a high-level semantic plan for translation. In contrast to gloss-based pipelines, these thought states are learned end-to-end under sentence-level supervision and are explicitly organized to form a directional latent chain. Given encoder features $\mathbf{E} = \{e_t\}_{t=1}^{T_s}$ and corresponding source mask, the module outputs $\mathbf{C} = \mathrm{Think}(\mathbf{E}) \in \mathbb{R}^{K \times d}$, where the index k defines the causal order of the latent reasoning process.

Learnable thought slots. A central difficulty in gloss-free SLT is that the video provides dense and redundant evidence, while the translation requires a compact sequence of semantic decisions. Instead of forcing the decoder to discover this structure on the fly, we introduce an explicit intermediate interface: a small set of ordered thought slots that serve as the model's working memory for step-wise semantic composition. Concretely, we initialize the thought chain with K learnable slots $\mathbf{C}^{(0)} = \{c_k^{(0)}\}_{k=1}^{K}$, where each $c_k^{(0)} \in \mathbb{R}^d$ is a trainable parameter, serving as empty reasoning states before evidence injection. These slots can be viewed as *empty* reasoning states that will be filled with semantic evidence distilled from $\mathbf{E}$.

Thinking layers. A key bottleneck in gloss-free SLT is that the encoder produces a long, continuous evidence stream $\mathbf{E}$, while what we ultimately need for generation is a compact and structured semantic plan. If we let the decoder directly attend to all frames, it must simultaneously decide what to express and where the supporting evidence lies, which often leads to diffuse attention and unstable evidence tracking. We therefore introduce a dedicated thinking stage that repeatedly compresses and reorganizes $\mathbf{E}$ into an ordered latent chain $\mathbf{C}$ before decoding. To make this compression process expressive yet controllable, we stack L identical thinking layers to iteratively refine the chain:

$$\mathbf{C}^{(\ell)} = \mathrm{ThinkLayer}^{(\ell)}(\mathbf{C}^{(\ell-1)}, \mathbf{E}), \ \ell = 1, \ldots, L, \tag{3}$$

and set $\mathbf{C} = \mathbf{C}^{(L)}$ as the final latent chain-of-thought. Each thinking layer consists of two main steps: causal thought mixing to impose an ordered chain prior, and an evidence-routed cross-attention operator that retrieves sign evidence in a structured manner.

We first encourage a directional (coarse-to-fine) refinement inside the latent chain. Specifically, we apply masked self-attention over the K thought states using a lower-triangular (causal) mask, so that the k-th thought only attends to $\{1, \ldots, k\}$:

$$\tilde{\mathbf{C}}^{(\ell)} = \mathrm{CausalSelfAttn}\Big(\mathbf{C}^{(\ell-1)}\Big). \tag{4}$$

This simple prior turns the thought set into an ordered workspace: earlier thoughts summarize coarse semantics and provide context, while later thoughts refine with additional details, making the intermediate reasoning process more stable and interpretable.

The second challenge is evidence assignment. A standard cross-attention from thoughts to all frames is both inefficient on long clips and prone to degeneracy, where multiple thoughts attend to the same frames. We address this by routing before retrieval: each thought is first encouraged to be approximately contiguous (via Lcont), and only then does it retrieve fine-grained evidence within that responsibility region.

From the frame-level stream $\mathbf{E}$, we derive (i) frame evidence (the original tokens), (ii) a small set of segment-level tokens $\mathbf{S}$ obtained by a differentiable soft segmentation over time (Chen

et al., 2025). Intuitively, $\mathbf{S}$ provides sparse semantic "chunks" (analogous to latent phrase units) while frame evidence preserves fine-grained lexical details. We compute segment-to-frame weights $\mathbf{W}_{seg} \in \mathbb{R}^{M \times T_s}$ via soft boundaries (per instance), where each row defines a soft temporal window and is normalized over frames. Segment tokens are obtained by weighted pooling:

$$S_j = \sum_{t=1}^{T_s} (\mathbf{W}_{seg})_{j,t}\, e_t, \qquad j = 1, \ldots, M. \quad (5)$$

Given segment tokens $\mathbf{S} = \{S_j\}_{j=1}^{M}$ and latent thoughts $\mathbf{C} = \{c_k\}_{k=1}^{K}$, we compute a similarity score matrix $\mathbf{G} \in \mathbb{R}^{K \times M}$ with

$$G_{k,j} = \frac{(W_q c_k)^\top (W_k S_j)}{\sqrt{d}}. \quad (6)$$

We then apply a differentiable Sinkhorn-style normalization to obtain a soft binding matrix $\mathbf{A} = \mathrm{Sinkhorn}(\mathbf{G}) \in \mathbb{R}^{K \times M}$, where $A_{k,j} \geq 0$ denotes the (soft) amount of segment j assigned to thought k. In practice, we perform Sinkhorn normalization with a *row-wise unit mass* and a *uniform column budget*, i.e., $\sum_{j=1}^{M} A_{k,j} \approx 1$ for all k and $\sum_{k=1}^{K} A_{k,j} \approx K/M$ for all j (up to a small error ϵ_{num}). This yields an approximately balanced transport plan: each thought receives comparable total evidence, while each segment contributes a limited and evenly distributed amount across thoughts, preventing collapsed routing while keeping the assignment fully soft. Each thought receives a routed segment summary:

$$p_k = \sum_{j=1}^{M} A_{k,j}\, S_j, \qquad p_k \in \mathbb{R}^d, \quad (7)$$

where p_k aggregates the segment evidence assigned to thought c_k. Since $\mathbf{A}$ is approximately row-normalized (i.e., $\sum_j A_{k,j} \approx 1$), p_k can be interpreted as a weighted average of segment tokens.

Conditioned on the routed summary p_k, each thought retrieves fine-grained sign evidence from $\mathbf{E}$ via cross-attention (and optionally sparse/deformable retrieval for efficiency on long clips). We denote the resulting evidence-aware thought update by

$$\hat{\mathbf{C}}^{(\ell)} = \mathrm{RoutedXAttn}\Big(\tilde{\mathbf{C}}^{(\ell)}, \mathbf{E}, \mathbf{p}\Big), \mathbf{p} = \{p_k\}_{k=1}^{K} \quad (8)$$

where RoutedXAttn encapsulates *(evidence-fabric → routing → retrieval)* as a single operator in our implementation and in the overview figure, and uses p_k (equivalently, the routing induced by $\mathbf{A}$) to bias the cross-attention over $\mathbf{E}$. Finally, we apply a position-wise feed-forward network with residual normalization to produce the layer output:

$$\mathbf{C}^{(\ell)} = \mathrm{FFN}\Big(\hat{\mathbf{C}}^{(\ell)}\Big). \quad (9)$$

The final thought chain $\mathbf{C}$ is consumed by the decoder through thinking cross-attention (Sec. 3.3). In addition, the intermediate routing variables (e.g., the thought-to-segment binding $\mathbf{A}$ and the segmentation weights) can be cached as an internal prior to guide visual grounding in the dual-stream decoder, without changing the model Input and Output.

3.3 Dual-Stream Decoder

We adopt an autoregressive Transformer decoder that explicitly follows a *plan-then-ground* generation pattern. At each layer, the decoder first consults the latent thought chain to decide *what* semantic content to express, and then returns to the encoder features to retrieve fine-grained signing evidence for *how* to realize that content in fluent text.

Given an input text prefix $\mathbf{y}_{<t}$ (teacher forcing during training), we embed tokens with standard word and positional (and optionally token-type) embeddings to obtain decoder states.

Let $\mathbf{H}^{(l-1)}$ denote the decoder hidden states entering layer l. Each layer contains the following sub-layers. We first apply masked self-attention over previously generated tokens to produce $\mathbf{H}_{\mathrm{self}}^{(l)} = \mathrm{MSA}\big(\mathbf{H}^{(l-1)}\big)$, where the subsequent mask ensures autoregressive generation.

We then attend to the latent thought chain $\mathbf{C}$:

$$\mathbf{H}_{\mathrm{think}}^{(l)} = \mathrm{XAttn}\Big(\mathbf{H}_{\mathrm{self}}^{(l)}, \mathbf{C}\Big). \quad (10)$$

Intuitively, this step selects the current semantic "reasoning state" from the ordered thoughts, providing a high-level plan for the next tokens. We also retain the corresponding attention weights over thoughts, denoted by $\boldsymbol{\alpha}$, as they reflect which thought(s) each target position relies on.

To make "think first, then ground" a mechanism rather than only an ordering, we derive a soft temporal retrieval prior from $\boldsymbol{\alpha}$ using intermediate routing variables produced by the thinking module (per instance). Specifically, we reuse (i) the thought-to-segment binding matrix $\mathbf{A} \in \mathbb{R}^{K \times M}$, and (ii) the segment-to-frame aggregation weights

$\mathbf{W}_{seg} \in \mathbb{R}^{M \times T_s}$. We map token-to-thought attention to a token-to-frame prior:

$$\boldsymbol{\beta} = \boldsymbol{\alpha}\mathbf{A}, \qquad \mathbf{w} = \boldsymbol{\beta}\mathbf{W}_{seg}, \tag{11}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{T_t \times K}$ denotes the head-averaged *attention weights* over K thoughts for each target position, and $\mathbf{w} \in \mathbb{R}^{T_t \times T_s}$ assigns each target position a soft preference over source time steps.

Here $\mathbf{A}$ is approximately row-normalized over segments (i.e., $\sum_{j=1}^{M} A_{k,j} \approx 1$) and $\mathbf{W}_{seg}$ is row-normalized over frames (i.e., $\sum_{t=1}^{T_s} (\mathbf{W}_{seg})_{j,t} = 1$), with all entries non-negative, so that each row of $\mathbf{w}$ forms an (approximately) normalized soft temporal prior. This routing prior is internal (no change to model I/O) and can be omitted without affecting the decoder interface. Finally, we attend to the encoder features $\mathbf{E}$ to obtain grounded decoder states:

$$\mathbf{H}_{\text{vis}}^{(l)} = \text{XAttn}\Big(\mathbf{H}_{\text{think}}^{(l)}, \mathbf{E}\Big), \tag{12}$$

with the standard source mask applied. When the routing prior $\mathbf{w}$ is enabled, we incorporate it as a soft additive bias in the attention logits, encouraging the decoder to retrieve evidence from time regions consistent with the selected thought(s). This yields thought-guided evidence retrieval while preserving the standard Transformer decoder structure (Li et al., 2026a; Hu et al., 2026).

A position-wise FFN with residual connections and layer normalization produces the layer output: $\mathbf{H}^{(l)} = \text{FFN}\Big(\mathbf{H}_{\text{vis}}^{(l)}\Big)$. The final decoder states are projected to the vocabulary to obtain token probabilities, and the model is trained with the standard sequence-level cross-entropy objective (Sec. 3.4).

3.4 Training Objective

Translation loss. Given a mini-batch of paired data $\{(\mathbf{X}^{(b)}, \mathbf{Y}^{(b)})\}_{b=1}^{B}$, where B is the batch size and $T_t^{(b)}$ denotes the target length of the b-th sample, we train the dual-stream decoder with teacher forcing and standard cross-entropy:

$$\mathcal{L}_{ce} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{T_t^{(b)}} \log p_\theta\Big(y_t^{(b)} \mid y_{<t}^{(b)}, \mathbf{E}^{(b)}, \mathbf{C}^{(b)}\Big). \tag{13}$$

Latent chain regularization. To encourage the latent thoughts to behave as an ordered chain of contiguous semantic units, we add two simple regularizers on the thought-to-segment assignment matrix $\mathbf{A}$ produced by the thinking module (Sec. 3.2).

Let μ_k be the expected segment index assigned to the k-th thought:

$$\mu_k^{(b)} = \sum_{j=1}^{M} j \cdot A_{k,j}^{(b)} \tag{14}$$

We penalize violations of the forward (coarse-to-fine) progression:

$$\mathcal{L}_{mono} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K-1} \text{ReLU}(\mu_k^{(b)} - \mu_{k+1}^{(b)} + \delta) \tag{15}$$

where $\delta = 1$ is a small margin.

We encourage each thought to focus on a contiguous region in segment space using a total-variation penalty:

$$\mathcal{L}_{cont} = \frac{1}{BK}\sum_{b=1}^{B}\sum_{k=1}^{K}\sum_{j=2}^{M} |A_{k,j}^{(b)} - A_{k,j-1}^{(b)}| \tag{16}$$

Overall objective. Our final training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_{mono}\mathcal{L}_{mono} + \lambda_{cont}\mathcal{L}_{cont} \tag{17}$$

where $\lambda_{mono} = 0.1$ and $\lambda_{cont} = 0.2$ control the strength of structural regularization. All components (encoder, thinking module, and decoder) are trained end-to-end with sentence-level supervision.

4 LC-HKSLT Dataset

Despite recent progress on gloss-free SLT, existing benchmarks are often limited in scale or do not fully reflect realistic deployment, where only videos and sentence-level translations (without glosses or SLR vocabularies) are available and may be noisy. This setting places greater emphasis on robust cross-modal reasoning and grounding. To study large-scale, gloss-free SLT in a realistic setting, we construct **LC-HKSLT**, a Hong Kong Sign Language corpus curated from broadcast-style public briefings with continuously visible interpreters. It contains 1,311 hours (432K clips) and provides only sentence-level supervision (no glosses or SLR vocabularies), matching our intended deployment regime. Full details are provided in Appendix C.1.

5 Experiments

5.1 Experimental Setup

Datasets. Following prior works (Chen et al., 2022a; Zhou et al., 2023, 2024a), we conduct experiments on five SLT benchmarks: PHOENIX-2014T (DGS) (Camgoz et al., 2018), CSL-Daily (CSL) (Zhou et al., 2021), How2Sign

Method	Modality		PHOENIX14T					CSL-Daily				
	Pose	RGB	B@1	B@2	B@3	B@4	ROUGE	B@1	B@2	B@3	B@4	ROUGE
Gloss-based												
SLRT (Camgoz et al., 2020)		✔	46.61	33.73	26.19	21.32	-	37.38	24.36	16.55	11.79	36.74
SignBT (Zhou et al., 2021)		✔	50.80	37.75	29.72	24.32	49.54	51.42	37.26	27.76	21.34	49.31
MMTLB (Chen et al., 2022a)		✔	53.97	41.75	33.84	28.39	52.65	53.31	40.41	30.87	23.92	53.25
SLTUNET (Zhang et al., 2023)		✔	52.92	41.76	33.99	28.47	52.11	54.98	41.44	31.84	25.01	54.08
CV-SLT (Zhao et al., 2024)		✔	54.88	42.68	34.79	29.27	54.33	58.29	45.15	35.77	28.94	57.06
TS-SLT (Chen et al., 2022b)	✔	✔	54.90	42.43	34.46	28.95	53.48	55.44	42.59	32.87	25.79	55.72
Weakly supervised gloss-free												
GASLT (Yin et al., 2023)		✔	39.07	26.74	21.86	15.74	39.86	19.90	9.94	5.98	4.07	20.35
VAP (Jiao et al., 2024)	✔		53.07	-	-	26.16	51.28	52.98	-	-	23.65	51.09
Gloss-free												
NSLT (Camgoz et al., 2018)		✔	32.24	19.03	12.83	9.58	31.80	34.16	19.57	11.84	7.56	34.54
GFSLT-VLP (Zhou et al., 2023)		✔	43.71	33.18	26.11	21.44	42.29	39.37	24.93	16.26	11.00	36.44
MSLU (Zhou et al., 2024a)	✔		46.56	34.21	-	22.24	46.73	33.97	22.20	-	11.42	33.80
FLa-LLM (Chen et al., 2024b)		✔	46.29	35.33	28.03	23.09	45.27	37.13	25.12	18.38	14.20	37.25
Sign2GPT (Wong et al., 2024)		✔	49.54	35.96	28.83	22.52	48.90	41.75	28.73	20.60	15.40	42.36
SignLLM (Gong et al., 2024)		✔	45.21	34.78	28.05	23.40	44.49	39.55	28.13	20.07	15.75	39.91
C^2RL (Chen et al., 2024a)		✔	52.81	40.20	**32.20**	26.75	50.96	49.32	36.28	27.54	21.61	48.21
SignThought (Ours)		✔	51.18	39.40	32.17	**27.22**	**54.50**	**49.57**	**37.90**	**29.33**	**23.92**	**50.99**

Table 1: SLT results on PHOENIX14T and CSL-Daily datasets. The best results are highlighted as **bold**, and the second-best are underlined.

Method	Modality		How2Sign					OpenASL				
	Pose	RGB	B@1	B@2	B@3	B@4	ROUGE	B@1	B@2	B@3	B@4	ROUGE
GloFE-VN (Lin et al., 2023)	✔		14.94	7.27	3.93	2.24	12.61	21.56	12.74	9.05	7.06	21.75
T5-SLT (Uthus et al., 2023)	✔		14.96	5.11	2.26	1.22	-	-	-	-	-	-
T5-SLT-YT (Uthus et al., 2023)	✔		37.82	24.13	16.92	12.39	-	-	-	-	-	-
VAP (Jiao et al., 2024)	✔		**39.22**	-	-	12.87	27.77	**45.92**	-	-	**21.23**	**41.38**
I3D-transformer (Shi et al., 2022a)		✔	-	-	-	-	-	18.31	10.15	7.19	5.66	18.64
OpenASL (Shi et al., 2022a)		✔	-	-	-	-	-	20.92	12.08	8.59	6.72	21.02
TF-H2S (Alvarez et al.)		✔	17.40	7.69	3.97	2.21	-	-	-	-	-	-
SLT-IV (Tarrés et al., 2023)		✔	34.01	19.30	12.18	8.03	-	-	-	-	-	-
C^2RL (Chen et al., 2024a)		✔	29.07	18.56	12.92	9.37	27.02	31.46	21.85	16.58	13.21	31.36
SignThought (Ours)		✔	36.65	**20.96**	**16.27**	**13.39**	**27.85**	38.10	**26.33**	**23.12**	19.55	38.78

Table 2: How2Sign and OpenASL datasets. T5-SLT-YT denotes training on YouTube-ASL and then fine-tuning on the 30-hour split How2Sign. The best results are highlighted as **bold**, and the second-best are underlined.

Method	B@1	B@2	B@3	B@4	ROUGE
NSLT (Camgoz et al., 2018)	29.91	19.56	13.02	8.65	9.67
TSPNet (Li et al., 2020b)	31.36	23.19	17.27	12.83	38.37
GASLT (Yin et al., 2023)	38.65	28.27	24.05	19.40	44.92
GFSLT-VLP (Zhou et al., 2023)	37.37	29.73	23.03	19.64	45.05
SignThought (Ours)	**39.12**	**31.75**	**25.93**	**21.15**	**47.87**
SignThought (Ours)†	45.33	38.01	31.83	30.22	60.01

Table 3: SLT results on LC-HKSLT. We report scores for four publicly available open-source gloss-free baselines and our SignThought. † denotes a variant that is pre-trained on the remaining LC-HKSLT data and then fine-tuned on the 30-hour split.

(ASL) (Duarte et al., 2021), OpenASL (ASL) (Shi et al., 2022b), and our LC-HKSLT corpus.

Evaluation Metrics. In line with previous studies (Chen et al., 2022a,b; Zhou et al., 2023), we report BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) for SLT. BLEU-n measures n-gram precision; we present BLEU-1 to BLEU-4 (B1-B4). ROUGE-L computes the F1 score based on the longest common subsequence between the hypothesis and the reference.

5.2 Comparison with SOTA Methods

Tab. 1 and 2 compare SignThought with representative gloss-free SLT methods across five benchmarks. SignThought achieves the best gloss-free BLEU-4 on all datasets and attains the highest ROUGE on PHOENIX14T, How2Sign, OpenASL, and LC-HKSLT, while remaining competitive on CSL-Daily. On PHOENIX14T, SignThought reaches **27.22** BLEU-4 and 54.50 ROUGE, surpassing the strongest prior gloss-free method C^2RL (26.75 / 50.96) as well as recent LLM-assisted baselines. On CSL-Daily, it achieves **23.92** BLEU-4 and 50.99 ROUGE, improving over C^2RL and slightly outperforming VAP in BLEU-4. Notably, larger gains are observed on large-scale datasets, with BLEU-4 improving from 9.37 to **13.39** on How2Sign and from 13.21 to **19.55** on OpenASL, accompanied by consistent ROUGE improvements. On LC-HKSLT, SignThought establishes a new

ID	Thk	Cau	Rtg	DDec	Pr	$\mathcal{L}_m$	$\mathcal{L}_c$	B4↑	R↑
0	✓	✓	✓	✓	✓	✓	✓	**27.49**	**55.90**
1	✗	–	–	✓	✗	✗	✗	25.30	51.20
2	✓	✗	✓	✓	✓	✓	✓	26.50	53.60
3	✓	✓	✗	✓	✓	✓	✓	26.10	53.00
4	✓	✓	✓	✗	✗	✓	✓	26.20	53.10
5	✓	✓	✓	✓	✗	✓	✓	26.60	53.90
6	✓	✓	✓	✓	✓	✗	✓	26.70	54.00
7	✓	✓	✓	✓	✓	✓	✗	26.75	54.05
8	✓	✓	✓	✓	✓	✗	✗	26.20	53.20

Table 4: Key element ablation in SignThought. Thk: latent thinking module; Cau: causal self-attn on thoughts; Rtg: structured routing; DDec: dual-stream decoder; Pr: prior injection; $\mathcal{L}_m/\mathcal{L}_c$: monotonicity/contiguity regularizers. "–" denotes not applicable.

state of the art among publicly available methods, achieving **21.15** BLEU-4 and **47.87** ROUGE. Moreover, due to the limited scale of existing public SLT benchmarks, most prior methods cannot perform additional in-domain pre-training under comparable settings. By pre-training SignThought on the remaining LC-HKSLT data and fine-tuning on the curated 30-hour split, we obtain substantial additional gains (denoted by †), highlighting the critical role of scaling in-domain sign–text data.

5.3 Ablation Study

Impact of Key Components in SignThought We conduct ablation studies on the PHOENIX14T dev set (Tab. 4); additional results are reported in the Appendix. Removing the latent thinking module (ID 1) leads to the largest performance drop, confirming the necessity of an explicit intermediate reasoning interface. Disabling causal thought updates (ID 2) further degrades accuracy, indicating that chain-structured reasoning is superior to unordered latent representations. Replacing structured routing with soft routing (ID 3) and simplifying the dual-stream decoder to a single stream (ID 4) both impair performance, highlighting the importance of explicit evidence allocation and plan-then-ground decoding. Removing thought-guided prior injection (ID 5) results in a smaller yet consistent decline, suggesting its role in emphasizing relevant temporal evidence. Finally, ablations of $\mathcal{L}mono$ and $\mathcal{L}cont$ (IDs 6–8) yield limited individual impact but a pronounced drop when jointly removed, implying complementary effects in stabilizing coherent thought-to-evidence alignment.

Hyperparameter and Structure-Strength Ablations. On PHOENIX14T, we vary one factor at a time while keeping the rest of the full model unchanged. We sweep the number of thoughts $K \in \{2, 4, 6, 8, 10\}$ (reasoning slots), the thinking depth $L \in \{1, 2, 3, 4, 5\}$ (refinement steps), and the number of segment tokens $M \in \{4, 8, 16, 32, 64\}$ (evidence granularity for routing) to study the trade-off between reasoning capacity and routing complexity and to locate a practical sweet spot.

Vary K fix L=2, M=16			**Vary** L fix K=8, M=16		
K	B4↑	R↑	L	B4↑	R↑
2	25.90	52.00	1	26.70	53.90
4	26.60	53.10	2	**27.49**	**55.90**
6	27.00	54.10	3	27.18	54.40
8	**27.49**	**55.90**	4	27.05	54.10
10	27.10	54.30	5	26.73	54.02

Vary M fix K=8, L=2		
M	B4↑	R↑
4	26.50	53.30
8	26.95	54.00
16	**27.49**	**55.90**
32	27.12	54.30
64	26.80	53.70

Table 5: Hyperparameter/structure-strength ablations.

Reference:	我完全看不出这两张照片有什么差别。 (I totally can't tell what the difference is between these two photos.)
GFSLT-VLP:	我觉得这照片有问题。 (I think there's something wrong with this photo.)
SignThought:	我真的看不出这两张照片有什么区别。 (I really can't tell what the distinction between these two photos.)
Reference:	这场大雨已经下了一整天了。 (This heavy rain has been falling all day.)
GFSLT-VLP:	外面正在下雨一天。 (It's raining outside for one day.)
SignThought:	这场大雨下了整整一天。 (This heavy rain has been falling for the whole day.)
Reference:	晚上我请你去吃肯德基。 (I'll treat you to KFC tonight.)
GFSLT-VLP:	晚上一起去吃饭吧。 (Let's go out for dinner tonight.)
SignThought:	今晚我请你去吃肯德基。 (Tonight, I'll treat you to KFC.)
Reference:	走在一对情侣旁边，我感觉自己很多余。 (Walking next to a couple, I feel like a third wheel.)
GFSLT-VLP:	我走在他们之间，感觉不太自在。 (Walking between them, I don't feel very comfortable.)
SignThought:	走在那对情侣旁边，我觉得自己特别多余。 (Walking next to that couple, I feel like such a third wheel.)

Table 6: Translation results. Correct answers are marked in green, semantically similar but rephrased answers are displayed in blue, and incorrect answers are indicated in red.

5.4 Qualitative Results

Table 6 presents representative CSL-Daily cases comparing SignThought with GFSLT-VLP, a only recent open-source baseline for gloss-free SLT. Following the color scheme in the caption, SignThought consistently preserves the core semantics of the reference while allowing minor paraphrases (blue), whereas GFSLT-VLP more often exhibits semantic drift or missing key information (red), e.g., over-generalizing "difference between two photos" to a vague "something wrong," weakening duration expressions such as "all day," or dropping salient entities (e.g., *KFC*). These examples qualitatively support our claim that the proposed latent reasoning and evidence allocation help maintain faithful, grounded translations, especially for sentences requiring fine-grained semantic composition.

6 Conclusion

We propose SignThought, a reasoning-driven framework for gloss-free sign language translation that introduces Cross-Modal Latent Thoughts and a Latent Chain-of-Thought to bridge continuous sign videos and text generation without gloss supervision. SignThought achieves SOTA performance across five benchmarks, demonstrating that latent-thought reasoning provides a scalable alternative to gloss annotations. We further introduce LC-HKSLT, a large-scale Cantonese SLT dataset to facilitate evaluation in realistic settings. Future work will explore stronger reasoning supervision, improved training and inference efficiency, and extensions to broader sign languages and open-world scenarios.

Limitation

Although SignThought introduces an ordered thought chain, the "thinking" process in our framework remains latent rather than explicit. The intermediate thoughts are continuous hidden states that are only indirectly learned from the final translation objective, rather than being verbalized, externally supervised, or exposed as human-interpretable reasoning steps. As a result, while the model benefits from improved planning and grounding, its intermediate reasoning is still difficult to directly inspect, verify, or control. In particular, we cannot guarantee that each latent thought corresponds to a stable semantic concept or a human-recognizable reasoning unit, and error analysis remains largely outcome-based at the final translation level. Therefore, the current framework should be viewed as a step toward reasoning-aware sign language translation, rather than a system that already produces explicit and fully interpretable reasoning traces. An important direction for future work is to bridge latent planning with more explicit forms of reasoning, such as textual rationales, gloss-like abstractions, or controllable semantic plans (Dong et al., 2026).

Acknowledgement

The research described in this paper was supported by the National Natural Science Foundation of China (Grant No. 62372314). This work was also supported by computational resources provided by The Centre for Large AI Models (CLAIM) of The Hong Kong Polytechnic University.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24:1750–1762.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer.

Patricia Cabot Alvarez, Xavier Giró Nieto, and Laia Tarrés Benet. Sign language translation based on transformers for the how2sign dataset.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, and Tessa Verhoef. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 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.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5120–5130.

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

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. 2024a. C^2rl: Content and context representation learning for gloss-free sign language translation and retrieval.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024b.

Factorized learning assisted with large language model for gloss-free sign language translation. pages 7071–7081.

Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. 2026. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20463–20471.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. 2025. Offset: Segmentation-based focus shift revision for composed image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, page 6113–6122.

Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. 2020. Fully convolutional networks for continuous sign language recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 697–714. Springer.

Kearsy Cormier, David Quinto-Pozos, Zed Sevcikova, and Adam Schembri. 2012. Lexicalisation and delexicalisation processes in sign languages: Comparing depicting constructions and viewpoint gestures. *Language & communication*, 32(4):329–348.

Haonan Dong, Kehan Jiang, Haoran Ye, Wenhao Zhu, Zhaolu Kang, and Guojie Song. 2026. Neureasoner: Towards explainable, controllable, and unified reasoning via mixture-of-neurons. *Preprint*, arXiv:2604.02972.

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. *Preprint*, arXiv:2008.08143.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916.

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

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. 2026. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Jiani Huang, Shijie Wang, Liangbo Ning, Wenqi Fan, Shuaiqiang Wang, Dawei Yin, and Qing Li. 2026. Towards next-generation recommender systems: A benchmark for personalized recommendation assistant with llms. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 217–226.

Jiani Huang, Xingchen Zou, Lianghao Xia, and Qing Li. 2025. Mr. rec: Synergizing memory and reasoning for personalized recommendation assistant with llms. *arXiv preprint arXiv:2510.14629*.

Kehan Jiang, Haonan Dong, Zhaolu Kang, Zhengzhou Zhu, and Guojie Song. 2026a. Foe: Forest of errors makes the first solution the best in large reasoning models. *Preprint*, arXiv:2604.02967.

Yiyang Jiang, Guangwu Qian, Jiaxin Wu, Qi Huang, Qing Li, Yongkang Wu, and Xiao-Yong Wei. 2026b. Self-paced learning for images of antinuclear antibodies. *IEEE Transactions on Medical Imaging*, 45(4):1661–1672.

Yiyang Jiang, Wengyu Zhang, Xulu Zhang, Xiao-Yong Wei, Chang Wen Chen, and Qing Li. 2024. Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 7249–7258, New York, NY, USA. Association for Computing Machinery.

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

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3416–3424.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

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

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*.

Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. 2026a. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23373–23381.

Zixu Li, Yupeng Hu, Zhiwei Chen, Shiqi Zhang, Qinlei Huang, Zhiheng Fu, and Yinwei Wei. 2026b. Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6762–6770.

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

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. *arXiv preprint arXiv:2305.12876*.

Peiyang Liu, Ziqiang Cui, Di Liang, and Wei Ye. 2025a. Who stole your data? a method for detecting unauthorized rag theft. *arXiv preprint arXiv:2510.07728*.

Peiyang Liu, Xi Wang, Ziqiang Cui, and Wei Ye. 2025b. Queries are not alone: Clustering text embeddings for video search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 874–883.

Peiyang Liu, Jinyu Yang, Lin Wang, Sen Wang, Yunlai Hao, and Huihui Bai. 2023. Retrieval-based unsupervised noisy label detection on text data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4099–4104.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Chong-Wah Ngo, Yu-Gang Jiang, Xiao-Yong Wei, Wanlei Zhao, Feng Wang, Xiao Wu, and Hung-Khoon Tan. 2008. Beyond semantic search: What you observe may not be what you think. In *IEEE Computer Society*.

Shuo Ni, Di Wang, He Chen, Haonan Guo, Ning Zhang, and Jing Zhang. 2025. Unigeoseg: Towards unified open-world segmentation for geospatial scenes. *arXiv preprint arXiv:2511.23332*.

Zhe Niu, Ronglai Zuo, Brian Mak, and Fangyun Wei. 2024. A hong kong sign language corpus collected from sign-interpreted tv news. *Preprint*, arXiv:2405.00980.

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.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022a. Open-domain sign language translation learned from online video. In *EMNLP*.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022b. Open-domain sign language translation learned from online video. *Preprint*, arXiv:2205.12870.

Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5634.

Guobin Tu and Di Weng. 2026. Easlt: Emotion-aware sign language translation. *Preprint*, arXiv:2601.03549.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Preprint*, arXiv:2306.15162.

Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. 2021. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *ICLR*.

Zhen-Qun Yang Xiao-Yong Wei. 2013. Coaching the exploration and exploitation in active learning for interactive video retrieval. *IEEE Transactions on Image Processing*, 22(3):955–968.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. 2025. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*.

Kevin Xu and Issei Sato. 2025. A formal comparison between chain-of-thought and latent thought. *Preprint*, arXiv:2509.25239.

Qianyun Yang, Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, and Liqiang Nie. 2026. Stable: Efficient hybrid nearest neighbor search via magnitude-uniformity and cardinality-robustness. *arXiv preprint arXiv:2604.01617*.

Jiayu Yao, He Chen, Yizhuang Xie, Ning Zhang, Mingxu Yang, and Liang Chen. 2025. S 2 net: Spatial-aligned and semantic-discriminative network for remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

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

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. 2025. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*.

Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A simple unified model for sign language translation. In *ICLR*.

Ning Zhang, Shuo Ni, Liang Chen, Tong Wang, and He Chen. 2025a. High-throughput and energy-efficient fpga-based accelerator for all adder neural networks. *IEEE Internet of Things Journal*.

Wengyu Zhang, Qi Tian, Yi Cao, Wenqi Fan, Dongmei Jiang, Yaowei Wang, Qing Li, and Xiao-Yong Wei. 2025b. Graphatc: advancing multilevel and multi-label anatomical therapeutic chemical classification via atom-level graph learning. *Briefings in bioinformatics*, 26(2):bbaf194.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *AAAI*, pages 19643–19651.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1316–1325.

Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. 2024a. Scaling up multimodal pre-training for sign language understanding. *arXiv preprint arXiv:2408.08544*.

Xiaoling Zhou, Wei Ye, Zhemg Lee, Lei Zou, and Shikun Zhang. 2025. Valuing training data via causal inference for in-context learning. *IEEE Transactions on Knowledge and Data Engineering*.

Xiaoling Zhou, Wei Ye, Yidong Wang, Chaoya Jiang, Zhemg Lee, Rui Xie, and Shikun Zhang. 2024b. Enhancing in-context learning via implicit demonstration augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2810–2828.

Inge Zwitserlood. 2012. Classifiers. In *Sign language: An international handbook*. de Gruyter.

A Method and Implementation Details

A.1 Implementation Details.

All experiments are conducted on a single NVIDIA RTX A6000 GPU. We implement our model in PyTorch and train all components (sign encoder, Latent CoT thinking module, and dual-stream decoder) end-to-end for 200 epochs. Unless otherwise specified, we use a shared model dimension d=256 throughout the encoder, thinking module, and decoder. All datasets use the same pre-extracted visual features of dimension 1024, following the feature extraction pipeline used in (Camgoz et al., 2020; Voskou et al., 2021), which adopts an identical approach with an open-sourced pretrained Inception network. The sign encoder consists of $N_{\text{enc}} = 2$ stacked encoder layers, and the Latent CoT thinking module contains $L = 2$ thinking layers with $K = 8$ learnable thought slots. The decoder is a standard autoregressive Transformer with $N_{dec} = 2$ layers, where each layer performs masked self-attention, thinking cross-attention to the thought chain, and visual cross-attention to encoder features (Sec. 3.3).

For the optional thought-guided evidence routing, we reuse the cached routing variables from the thinking module and apply them as a soft bias in the decoder's visual cross-attention. We initialize linear/embedding weights with Xavier initialization and apply dropout with a rate of 0.1 to attention and feed-forward sublayers. We use layer normalization in all Transformer-style blocks. We optimize with Adam (lr $= 1 \times 10^{-3}$, β_1=0.9, β_2=0.998) and a batch size of 32. We train the model with weight decay $= 3 \times 10^{-3}$ and label smoothing $= 0.1$. We use a plateau-based learning rate scheduler and apply a warmup phase of 2000 steps before the scheduler takes effect.

We evaluate on the validation set periodically and reduce the learning rate by a factor of 0.8 if the validation metric does not improve for several consecutive checks, stopping when the learning rate drops below 1×10^{-4}. At test time, we use beam search for translation decoding and tune decoding hyperparameters on the development set. We sweep the beam size $b \in [1, 10]$ and length-penalty coefficient $a \in \{-1, 0, 1, 2, 3, 4, 5\}$ on the dev set; the best (b, a) is used for test-time decoding.

A.2 Soft Segmentation via Soft Boundaries

Given encoder frame features $\mathbf{E} = \{e_t\}_{t=1}^{T_s}$, we construct M segment tokens $\mathbf{S} = \{S_j\}_{j=1}^{M}$ by predicting *soft* temporal windows with learnable boundaries (per instance). Let $\bar{m}_t \in \{0, 1\}$ denote a multiplicative padding mask over source frames ($\bar{m}_t = 0$ for padded frames and $\bar{m}_t = 1$ otherwise). Define the valid length $L_{\text{vaild}} = \sum_{t=1}^{T_s} \bar{m}_t$ (we assume $L_{\text{vaild}} \geq 1$).

We first obtain a clip-level summary by masked mean pooling:

$$z = \frac{\sum_{t=1}^{T_s} \bar{m}_t \, e_t}{L_{\text{vaild}} + \epsilon_{\text{num}}} \in \mathbb{R}^d. \tag{18}$$

We then predict positive segment lengths $\rho \in \mathbb{R}^M$ with a boundary MLP:

$$\rho = \text{softplus}(\text{MLP}_{\text{bd}}(z)), \tag{19}$$

$$\pi = \frac{\rho}{\sum_{j=1}^{M} \rho_j + \epsilon_{\text{num}}}, \tag{20}$$

and convert proportions into cumulative boundaries in continuous time over valid frames:

$$\tau_0 = 1, \;\; \tau_j = 1 + (L_{\text{vaild}} - 1) \sum_{i=1}^{j} \pi_i, \quad j = 1, \ldots, M, \tag{21}$$

so $\tau_M \approx L_{\text{vaild}}$ and each instance can have different segment durations even when M is fixed.

Soft window membership and weights. Let $\hat{t} = \sum_{u=1}^{t} \bar{m}_u$ denote the valid-frame rank of position t. For each segment j, we define a soft window membership over discrete frames:

$$u_{j,t} = \sigma(\gamma(\hat{t} - \tau_{j-1})) - \sigma(\gamma(\hat{t} - \tau_j)), \;\; t = 1, \ldots, T_s, \tag{22}$$

where $\sigma(\cdot)$ is the sigmoid and $\gamma > 0$ controls boundary softness (we set $\gamma = 1$). We apply the padding mask and normalize over frames:

$$u_{j,t} \leftarrow \bar{m}_t \, u_{j,t}, \;\; (\mathbf{W}_{seg})_{j,t} = \frac{u_{j,t}}{\sum_{t'=1}^{T_s} u_{j,t'} + \epsilon_{\text{num}}}, \tag{23}$$

so each row of $\mathbf{W}_{seg} \in \mathbb{R}^{M \times T_s}$ sums to 1.

Segment tokens. Finally, segment tokens are obtained by weighted pooling:

$$S_j = \sum_{t=1}^{T_s} (\mathbf{W}_{seg})_{j,t} \, e_t, \qquad j = 1, \ldots, M. \tag{24}$$

We use $\epsilon_{\text{num}} = 10^{-6}$ throughout for numerical stability.

A.3 Injecting routed summaries p_k into RoutedXAttn

We implement conditioning on p_k by converting it into a routing-dependent bias over source time steps and adding it to the cross-attention logits. For each head $h \in \{1, \dots, H\}$ with head dimension $d_h = d/H$, we compute standard logits:

$$\ell_{k,t}^{(h)} = \frac{\left(W_Q^{(h)} \tilde{c}_k\right)^\top \left(W_K^{(h)} e_t\right)}{\sqrt{d_h}}, \tag{25}$$

and add a p_k-dependent bias before softmax:

$$\tilde{\ell}_{k,t}^{(h)} = \ell_{k,t}^{(h)} + \lambda_p \, b_{k,t}^{(h)}(p_k, \mathbf{E}), \tag{26}$$

where $\lambda_p \geq 0$ controls the bias strength (we set $\lambda_p = 1$ in all experiments).

Attention with hard padding mask. Let $m \in \{0, -\infty\}^{T_s}$ be the additive attention mask over the T_s source frames, where for each $t \in \{1, \dots, T_s\}$, $m_t = 0$ if frame t is valid (non-padded) and $m_t = -\infty$ otherwise; then:

$$a_{k,t}^{(h)} = \mathrm{softmax}_t\left(\tilde{\ell}_{k,t}^{(h)} + m_t\right), \tag{27}$$

$$o_k^{(h)} = \sum_{t=1}^{T_s} a_{k,t}^{(h)} \left(W_V^{(h)} e_t\right). \tag{28}$$

A minimal content-based bias. A simple differentiable instantiation is:

$$b_{k,t}^{(h)}(p_k, \mathbf{E}) = \frac{\left(W_P^{(h)} p_k\right)^\top \left(W_B^{(h)} e_t\right)}{\sqrt{d_h}}. \tag{29}$$

Optional routing-derived temporal prior. If segment-to-frame weights $\mathbf{W}_{seg} \in \mathbb{R}^{M \times T_s}$ are available (row-normalized over frames), we can derive a purely routing-based temporal prior:

$$\mathbf{r}_k = \mathbf{A}_{k,:} \mathbf{W}_{seg} \in \mathbb{R}^{1 \times T_s}, \tag{30}$$

$$r_{k,t} = \sum_{j=1}^{M} A_{k,j} (\mathbf{W}_{seg})_{j,t}. \tag{31}$$

We then inject it as an additive log-prior:

$$b_{k,t}^{(h)} \leftarrow b_{k,t}^{(h)} + \log(r_{k,t} + \epsilon_{\mathrm{num}}), \tag{32}$$

with $\epsilon_{\mathrm{num}} = 10^{-6}$.

A.4 Injecting the temporal prior w into grounding cross-attention

We compute a token-to-frame prior $\mathbf{w} \in \mathbb{R}^{T_t \times T_s}$ as:

$$\boldsymbol{\beta} = \boldsymbol{\alpha}\mathbf{A}, \qquad \mathbf{w} = \boldsymbol{\beta}\mathbf{W}_{seg}, \tag{33}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{T_t \times K}$ are head-averaged decoder-to-thought attention weights (row-normalized over K thoughts). Under non-negativity and approximate row normalization of $\boldsymbol{\alpha}, \mathbf{A}, \mathbf{W}_{seg}$, each row $\mathbf{w}_{t,\cdot}$ is an (approximately) normalized soft prior over source frames.

For grounding cross-attention at layer l, logits are:

$$\ell_{t,s}^{(h)} = \frac{\left(W_Q^{(h)} h_{\mathrm{think},t}^{(l)}\right)^\top \left(W_K^{(h)} e_s\right)}{\sqrt{d_h}}. \tag{34}$$

When enabled, we add a soft bias:

$$\tilde{\ell}_{t,s}^{(h)} = \ell_{t,s}^{(h)} + \lambda_w \log(\mathbf{w}_{t,s} + \epsilon_{\mathrm{num}}), \tag{35}$$

where $\lambda_w \geq 0$ (we set $\lambda_w = 1$ in all experiments). We then apply the standard additive source mask $m_s \in \{0, -\infty\}$:

$$a_{t,s}^{(h)} = \mathrm{softmax}_s\left(\tilde{\ell}_{t,s}^{(h)} + m_s\right). \tag{36}$$

Let the K thought tokens at layer ℓ be $\tilde{\mathbf{C}}^{(\ell)} = [\tilde{c}_1, \dots, \tilde{c}_K]^\top \in \mathbb{R}^{K \times d}$. Let the encoder (frame-level) features be $\mathbf{E} = [e_1, \dots, e_{T_s}]^\top \in \mathbb{R}^{T_s \times d}$. Let the segment tokens be $\mathbf{S} = [S_1, \dots, S_M]^\top \in \mathbb{R}^{M \times d}$, and the Sinkhorn routing matrix be $\mathbf{A} \in \mathbb{R}^{K \times M}$ with non-negative entries and approximately row-normalized ($\sum_{j=1}^{M} A_{k,j} \approx 1$). The routed summary for thought k is

$$p_k = \sum_{j=1}^{M} A_{k,j} S_j, \qquad p_k \in \mathbb{R}^d, \tag{37}$$

and we denote $\mathbf{p} = \{p_k\}_{k=1}^K$.

Conditioned on p_k, RoutedXAttn performs cross-attention from the thought tokens to the encoder features, with a p_k-dependent bias that favors frames consistent with the routed segment content (or equivalently, the routing induced by $\mathbf{A}$).

Concretely, for each head $h \in \{1, \dots, H\}$ with head dimension $d_h = d/H$, we compute standard cross-attention logits

$$\ell_{k,t}^{(h)} = \frac{\left(W_Q^{(h)} \tilde{c}_k\right)^\top \left(W_K^{(h)} e_t\right)}{\sqrt{d_h}}, \tag{38}$$

and add a p_k-dependent bias term $b^{(h)}_{k,t}(p_k, \mathbf{E})$ before the softmax:

$$\tilde{\ell}^{(h)}_{k,t} = \ell^{(h)}_{k,t} + \lambda_p \, b^{(h)}_{k,t}(p_k, \mathbf{E}). \tag{39}$$

where $\lambda_p \geq 0$ controls the strength of the routed-summary bias (we set $\lambda_p = 1$ in all experiments unless otherwise specified). The attention weights and outputs follow the standard definition:

$$a^{(h)}_{k,t} = \mathrm{softmax}_t\Big(\tilde{\ell}^{(h)}_{k,t} + m_t\Big), \tag{40}$$

$$o^{(h)}_{k} = \sum_{t=1}^{T_s} a^{(h)}_{k,t} \left(W^{(h)}_V e_t\right), \tag{41}$$

where $\mathrm{softmax}_t$ denotes normalization over the source time index t. Here $m_t = -\infty$ for masked (padded) time steps and $m_t = 0$ otherwise; thus the mask remains a hard constraint.

A minimal instantiation is a content-based compatibility between the routed summary and each source frame:

$$b^{(h)}_{k,t}(p_k, \mathbf{E}) = \frac{\left(W^{(h)}_P p_k\right)^\top \left(W^{(h)}_B e_t\right)}{\sqrt{d_h}}, \tag{42}$$

which increases attention to frames whose features align with the routed segment summary.

Alternatively, when the segmentation weights $\mathbf{W}_{seg} \in \mathbb{R}^{M \times T_s}$ are available (row-normalized over frames), we can additionally derive a purely routing-based temporal prior

$$r_{k,t} = \sum_{j=1}^{M} A_{k,j} \, (\mathbf{W}_{seg})_{j,t}, \qquad t = 1, \ldots, T_s, \tag{43}$$

We denote the resulting temporal prior vector by $\mathbf{r}_k \in \mathbb{R}^{T_s}$ with entries $(\mathbf{r}_k)_t = r_{k,t}$. and inject it as a log-prior bias:

$$b^{(h)}_{k,t} \;\leftarrow\; b^{(h)}_{k,t} + \log(r_{k,t} + \epsilon_{\mathrm{num}}). \tag{44}$$

This realizes the "equivalently, the routing induced by $\mathbf{A}$" view: both p_k and $\mathbf{r}_k$ are functions of the same routing variables, with p_k summarizing *what* evidence is assigned, and $\mathbf{r}_k$ summarizing *where in time* it is expected to lie.

We write Eqs. (26)–(44) as

$$\hat{\mathbf{C}}^{(\ell)} = \mathrm{RoutedXAttn}\Big(\tilde{\mathbf{C}}^{(\ell)}, \mathbf{E}, \mathbf{p}\Big), \tag{45}$$

where $\mathrm{RoutedXAttn}$ performs standard multi-head cross-attention from thoughts to encoder features with the above biasing mechanism.

A.5 Injecting the temporal prior w into the Dual-Stream Decoder grounding cross-attention

Let T_t be the target length. We compute a token-to-frame prior $\mathbf{w} \in \mathbb{R}^{T_t \times T_s}$ per instance as

$$\boldsymbol{\beta} = \boldsymbol{\alpha}\mathbf{A}, \qquad \mathbf{w} = \boldsymbol{\beta}\mathbf{W}_{seg}, \tag{46}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{T_t \times K}$ is the head-averaged *attention weights* (row-normalized over K thoughts) from the decoder token positions to the K thought tokens. With non-negative $\boldsymbol{\alpha}, \mathbf{A}, \mathbf{W}_{seg}$ and approximate normalization (rows summing to ≈ 1), each row vector $\mathbf{w}_t \in \mathbb{R}^{T_s}$ (the t-th row of $\mathbf{w}$) forms an (approximately) normalized soft temporal prior over source time steps.

In the dual-stream decoder, we compute $\boldsymbol{\alpha}$ from the thought-attention sublayer first (standard attention producing weights), then form $\mathbf{w}$ via Eq. (46), and finally use $\mathbf{w}$ in the grounding (encoder) cross-attention sublayer as a bias. This ensures $\mathbf{w}$ is available before computing the grounded attention distribution.

Let $\mathbf{H}^{(l)}_{\mathrm{think}} \in \mathbb{R}^{T_t \times d}$ denote the decoder states entering the grounding cross-attention at layer l. Vanilla cross-attention to encoder features uses logits

$$\ell^{(h)}_{t,s} = \frac{\left(W^{(h)}_Q h^{(l)}_{\mathrm{think},t}\right)^\top \left(W^{(h)}_K e_s\right)}{\sqrt{d_h}}. \tag{47}$$

When the routing prior is enabled, we add a soft bias derived from $\mathbf{w}$ before the softmax:

$$\tilde{\ell}^{(h)}_{t,s} = \ell^{(h)}_{t,s} + \lambda_w \log(\mathbf{w}_{t,s} + \epsilon_{\mathrm{num}}), \tag{48}$$

where $\lambda_w \geq 0$ controls the strength of the temporal prior bias (we set $\lambda_w = 1$ in all experiments unless otherwise specified). Followed by the standard source mask m_s (padded frames remain excluded):

$$a^{(h)}_{t,s} = \mathrm{softmax}_s\Big(\tilde{\ell}^{(h)}_{t,s} + m_s\Big), \tag{49}$$

$$\mathbf{H}^{(l)}_{\mathrm{vis},t} = \mathrm{Concat}_h \left(\sum_{s=1}^{T_s} a^{(h)}_{t,s} W^{(h)}_V e_s\right) W_O. \tag{50}$$

where $\mathrm{softmax}_s$ denotes normalization over the source time index s. Thus $\mathbf{w}$ steers the grounding cross-attention toward time regions consistent with the routed thought mixture implied by $\boldsymbol{\alpha}$, while preserving the standard Transformer decoder structure. Setting $\lambda_w = 0$ disables the mechanism without changing the model I/O or the decoder interface.

B Additional Ablations

B.1 Ablations on the Evidence Fabric

To verify that the gains come from multi-granular evidence rather than extra computation, we ablate the evidence fabric while keeping the routing and the dual-stream decoder unchanged, and only vary which evidence tokens are provided (Table 7). We evaluate: (i) **Frame-only**, where the fabric consists of frame-level tokens $\mathbf{E}$ only; (ii) **Segment-only**, where the fabric consists of segment tokens $\mathbf{S}$ only; and (iii) **Global-only**, where we replace the fabric with a single pooled token g computed by masked mean pooling over $\mathbf{E}$ (used only for this ablation). Frame evidence preserves fine lexical cues but weakens long-range composition; segment evidence supports planning but loses fine details; global-only is the most compressed and thus the weakest.

Fabric Variant (PHOENIX14T)	B4↑	R↑
Frame-only (no seg)	26.40	53.40
Segment-only (no frame)	26.00	52.90
Global-only (single pooled token)	24.80	50.80
Frame+Segment (Full)	**27.49**	**55.90**

Table 7: Evidence fabric ablations on PHOENIX14T

B.2 Additional Analyses: Length Buckets and Interpretability Metrics

Length-bucket evaluation. To further examine whether SignThought improves *long-range semantic composition*, we evaluate BLEU-4 under target-length buckets on PHOENIX14T (Table 8). We divide the dev set into three equally sized groups by reference length $|\mathbf{y}|$ (in BPE tokens): SHORT ($|\mathbf{y}| \leq 9$), MEDIUM ($10 \leq |\mathbf{y}| \leq 17$), and LONG ($|\mathbf{y}| \geq 18$), corresponding to the lower, middle, and upper tertiles of the dev set length distribution.

As shown in Table 8, the improvements of SignThought over the baseline grow with target length: +0.4 for short sentences, +1.1 for medium, and +2.2 for long sentences. This trend suggests that the latent reasoning chain contributes most when translation requires *compositional aggregation of multiple visual evidences* or modeling long-range temporal dependencies, aligning with the intended "plan–then–ground" inductive bias.

Quantitative interpretability metrics. We additionally quantify whether the latent chain learns selective and ordered thought-to-evidence alignments. Metrics are computed from the final-layer thought-to-segment binding matrix $A \in \mathbb{R}^{K \times M}$ on the PHOENIX14T test set and averaged over samples. Lower values indicate more structured and interpretable alignments.

BLEU-4 by $\lvert\mathbf{y}\rvert$ buckets	Short	Medium	Long
w/o latent thinking	29.20	25.60	19.40
SignThought (Full)	29.60	26.70	21.60
Gain (Full – Base)	+0.40	+1.10	+2.20

Table 8: Length-bucket BLEU-4 on PHOENIX14T.

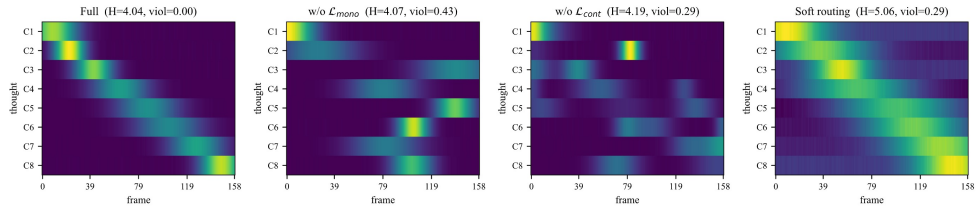
Method	Entropy↓	MonoViol↓	Span↓	TV↓
w/o latent thinking	2.10	0.27	18.5	0.34
SignThought (Full)	**1.45**	**0.12**	**14.0**	**0.22**

Table 9: Quantitative interpretability metrics on PHOENIX14T.

- **Entropy (↓)**: measures selectivity of each thought k as $H_k = -\sum_{j=1}^{M} A_{k,j} \log(A_{k,j} + \epsilon)$, averaged over k. Lower entropy indicates sharper attention (less diffuse binding).
- **Monotonicity violation (MonoViol, ↓)**: captures ordering consistency. Let $\mu_k = \sum_j j A_{k,j}$ be the expected segment index for thought k. We report the fraction of violations $\text{MonoViol} = \frac{1}{K-1} \sum_{k=1}^{K-1} \mathbb{I}[\mu_k > \mu_{k+1}]$, where smaller values mean that later thoughts attend to later evidence more consistently.
- **Span (↓)**: quantifies localization sharpness. For each k, define coverage indices $j_k^{(p)}$ satisfying $\sum_{j \leq j_k^{(p)}} A_{k,j} \geq p$, and compute $\text{Span}_k = j_k^{(0.95)} - j_k^{(0.05)}$. Smaller spans imply tighter evidence localization.
- **Total variation (TV, ↓)**: measures fragmentation of attention along time, $\text{TV} = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=2}^{M} |A_{k,j} - A_{k,j-1}|$. Lower TV indicates smoother, more contiguous alignments.

Table 9 reports these metrics, showing consistent improvements across all four aspects. SignThought exhibits lower entropy (1.45 vs. 2.10), fewer ordering violations (0.12 vs. 0.27), tighter localization (14.0 vs. 18.5), and smoother attention (0.22 vs. 0.34). Together, these results quantitatively confirm that the latent thought chain promotes coherent, ordered, and interpretable evidence allocation, aligning with the qualitative visualizations in Fig. 2.

Reference: 我完全看不出这两张照片有什么差别。(I totally can't tell what the difference is between these two photos.)



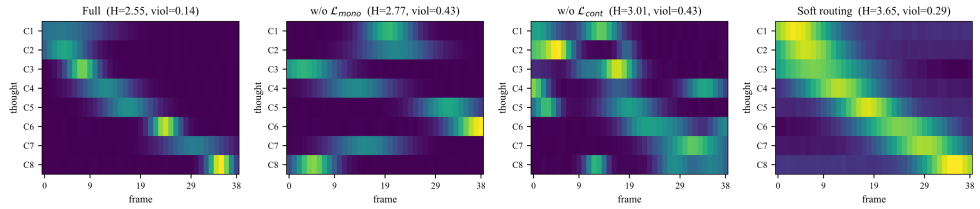
我真的看不出这两张照片有什么区别。(I really can't tell what the difference between these two photos.)

我看得出这两张照片有点差别。(I can tell these two photos are a bit different..)

我看不出这两张照片。有什么差别。(I can't tell these two photos. What difference they have.)

我觉得这两张照片不太一样。(I feel these two photos are not quite the same.)

Reference: 这场大雨已经下了一整天了。(This heavy rain has been falling all day.)



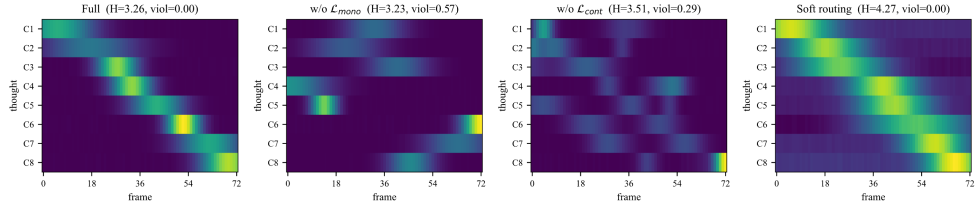
这场大雨已经下了一整天。(This heavy rain has been going on all day.)

这场大雨刚下了一会儿。(This heavy rain has only been falling for a short while.)

这场雨已经下了…一整天了。(This rain has been falling… all day.)

外面在下雨。(It's raining outside.)

Reference: 晚上我请你去吃肯德基。(I'll treat you to KFC tonight.)



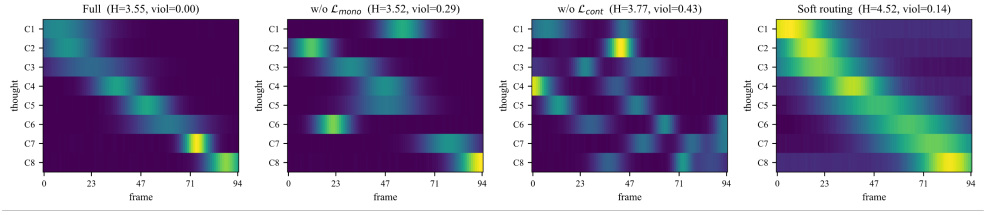
今晚我请你去吃肯德基。(Tonight, I'll treat you to KFC.)

今晚你请我去吃肯德基。(Tonight, you'll treat me to KFC.)

今晚请你去吃肯德基。(Tonight, I invite you to go to KFC.)

晚上去吃火锅。(Let's have hotpot tonight.)

Reference: 走在一对情侣旁边，我感觉自己很多余。(Walking next to a couple, I feel like a third wheel.)



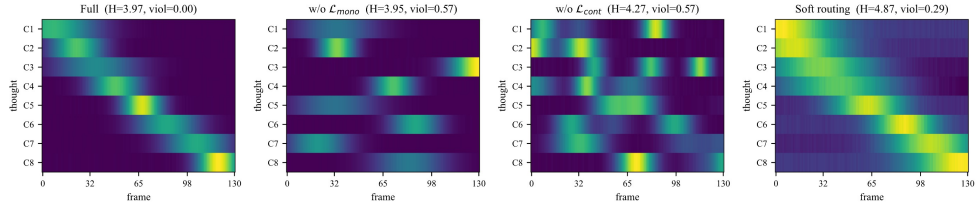
走在那对情侣旁边，我觉得自己特别多余。(Walking next to that couple, I feel like such a third wheel.)

走在一对情侣旁边，我觉得自己很好。(Walking next to a couple, I feel very good.)

我感觉自己不好走在路上。(I feel uneasy walking on the road..)

我走在旁边，感觉不太自在。(I walked beside them, feeling a bit uncomfortable.)

Reference: 不要担心，伯伯无论如何都要给你筹到学费。(Don't worry, your uncle will find a way to raise your tuition, no matter what.)



别担心，伯伯无论如何都会帮你筹到学费。(Do not worry, your uncle will help raise your tuition no matter what.)

不用担心，伯伯给学费。(No need to worry, your uncle give your tuition.)

担心，伯伯会给到学费。(Worry, your uncle will give tuition.)

别担心，总会有办法的。(No worry, there will be a way.)

Figure 3: Qualitative visualization on CSL-Daily.

B.3 Effect of Structural Regularization Strength

We study the sensitivity of the structured routing regularization to the two loss weights λ_{mono} and λ_{cont} in

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_{\text{mono}}\mathcal{L}_{mono} + \lambda_{\text{cont}}\mathcal{L}_{cont}. \quad (51)$$

These terms encourage (i) monotonic progression of the thought-to-segment assignment and (ii) reduced fragmentation (contiguity) in segment space, respectively. Concretely, we sweep $\lambda_{\text{mono}} \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$ and $\lambda_{\text{cont}} \in \{0, 0.05, 0.1, 0.2, 0.4\}$ while keeping all other settings fixed. For clarity, Table 10 reports representative one-dimensional sweeps on the development set: (i) varying λ_{mono} with λ_{cont} fixed to 0.2, and (ii) varying λ_{cont} with λ_{mono} fixed to 0.1.

Overall, introducing small-to-moderate structural regularization consistently improves translation quality over the no-regularization baseline ($\lambda_{\text{mono}} = \lambda_{\text{cont}} = 0$). Performance peaks at $\lambda_{\text{mono}} = 0.1$ with $\lambda_{\text{cont}} = 0.2$, while further increasing either coefficient degrades performance, suggesting over-regularization. Based on this sweep, we set $\lambda_{\text{mono}} = 0.1$ and $\lambda_{\text{cont}} = 0.2$ as the default choice in all experiments.

Sweep λ_{mono}			Sweep λ_{cont}		
λ_{mono}	λ_{cont}	Dev B4↑	λ_{mono}	λ_{cont}	Dev B4↑
0	0.2	26.10	0.1	0	27.10
0.01	0.2	26.55	0.1	0.05	27.35
0.05	0.2	27.10	0.1	0.1	27.22
0.1	**0.2**	**27.49**	**0.1**	**0.2**	**27.49**
0.2	0.2	27.25	0.1	0.4	26.64
0.5	0.2	26.50			
1.0	0.2	25.95			

Table 10: Development-set sensitivity to the structural regularization weights. We report one-dimensional sweeps by varying one coefficient and fixing the other to its selected default. The best setting is $\lambda_{\text{mono}} = 0.1$ and $\lambda_{\text{cont}} = 0.2$ (bold).

C More Details on LC-HKSLT: Scaling, Quality, and Ethics

C.1 LC-HKSLT Dataset

Our objective is to enable scalable, gloss-free sign language translation suitable for real-world deployment, where sign–text pairs are readily available but token-level annotations, such as glosses, are

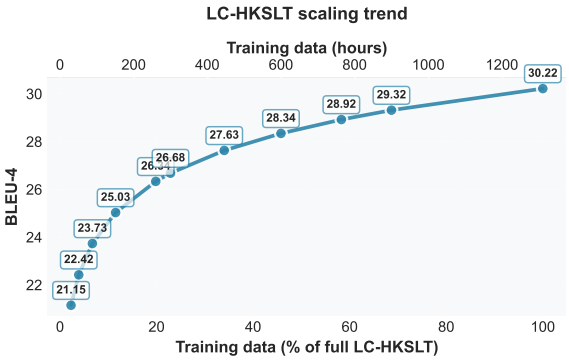


Figure 4: LC-HKSLT scaling trend.

scarce. Existing Hong Kong Sign Language resources are typically limited in scale or rely heavily on gloss supervision, which is misaligned with learning directly from raw videos and sentence-level translations. To bridge this gap and support our Latent Chain-of-Thought framework for long-range, weakly supervised sign-to-text reasoning, we construct LC-HKSLT, a large-scale corpus curated from realistic broadcast-style scenarios.

Collection pipeline. We collect approximately 1,300 hours of Hong Kong Sign Language videos from YouTube, primarily drawn from Hong Kong Government and Legislative Council briefings in which an interpreter is continuously visible and the spoken content is publicly accessible. The videos are segmented into sentence-level clips, from which frames are extracted, and the corresponding speech tracks are transcribed using `openai/whisper-large-v3` to produce sentence-level targets. Notably, no gloss annotations or sign language recognition vocabularies are introduced. The resulting supervision consists solely of weak, sentence-level signals that are readily obtainable at scale in real-world settings, thereby aligning with the intended operating regime of our gloss-free formulation.

Evaluation protocol. Although LC-HKSLT is collected at web scale, this paper focuses on a carefully curated 30-hour subset, with zero sentence overlap across the training, validation, and test splits. This design choice is not intended to underrepresent the dataset, but rather to isolate the methodological contribution in a controlled, comparable setting. Specifically, (i) most existing Chinese SLT benchmarks operate in the 20 to 30 hour regime, enabling fair and direct comparison; (ii) the reduced scale allows training to remain practical on a single GPU; and (iii) we observe that, even without external pre-

Dataset / Work	Lang.	SLR Vocab.	SLT Vocab.	Duration (h)	Videos	Signers	Source
GSL (Adaloglou et al., 2021)	GSL	310	–	9.59	10K	7	Lab
PHOENIX–14T (Camgoz et al., 2018)	DGS	1,066	2,887	10.53	8K	9	TV
CSL–Daily (Zhou et al., 2021)	CSL	2,000	2,370	23.27	21K	10	Lab
TVB–HKSL–News (Niu et al., 2024)	HKSL	6,515	2,850	16.07	7K	2	TV
LC–HKSLT (Ours)	HKSL	–	125,833	1311	432K	14	YouTube

Table 11: Comparison of LC–HKSLT with representative SLR/SLT datasets.

training, this data budget is sufficient to validate the core benefits of explicit latent reasoning. Table 11 situates LC-HKSLT among representative SLR and SLT datasets. Beyond the subset used in this study, the complete collection comprises 432k clips from 14 signers and covers 125,833 Chinese words and phrases in the translation vocabulary. The full dataset will be publicly released to support future research on large-scale, gloss-free SLT. Beyond direct translation, large-scale sign-text corpora may also support retrieval-oriented multimodal settings that require composing linguistic intent with fine-grained visual evidence (Li et al., 2026b).

C.2 Scaling Study on LC-HKSLT

To address concerns about the gap between our web-scale collection and the controlled 30-hour evaluation protocol, we conduct an additional scaling study on LC-HKSLT by training the same model with increasing fractions of the training pool while keeping the dev/test split fixed. As shown in Fig. 4, performance improves monotonically with more training data: BLEU-4 increases from 21.15 in the smallest-data setting to 30.22 when using the full LC-HKSLT collection (up to $\sim$1.3k hours), yielding a total gain of +9.07 BLEU-4. Notably, improvements are most pronounced in the low-data regime and gradually taper off as the data scale grows, suggesting that our method is data-efficient while continuing to benefit from additional weakly supervised sign-text pairs.

C.3 Data Quality, Licensing, and Privacy Considerations

Data source and intended use. LC-HKSLT is collected from publicly accessible broadcast-style videos on YouTube, focusing on Hong Kong Government and Legislative Council (LegCo) briefings where a sign language interpreter is consistently visible and the spoken content is publicly available. We segment videos into sentence-level clips and obtain sentence targets by transcribing the accompanying speech track using `openai/whisper-large-v3`. The dataset is intended solely for **non-commercial research and education** on sign language translation and related multimodal learning problems. More broadly, this line of work is also related to structured prediction over complex inputs, where multi-level semantic dependencies are important (Zhang et al., 2025b).

Label noise and alignment limitations. Since sentence targets are derived from automatic speech recognition (ASR), LC-HKSLT may contain: (i) ASR transcription errors, (ii) sentence boundary/segmentation errors, and (iii) imperfect temporal alignment between the interpreter's signing and the spoken content. Such issues are related to the broader challenge of learning under noisy supervision in text-centered settings (Liu et al., 2023, 2025b; Chen et al., 2026). We explicitly view LC-HKSLT as weak, sentence-level supervision at scale, which matches the real-world operating condition targeted by gloss-free SLT. We will release meta-information (e.g., clip timestamps and source video identifiers) to facilitate error analysis and downstream filtering by the community. Efficient hybrid nearest-neighbor search may further support scalable indexing and community-driven filtering over large clip collections (Yang et al., 2026; Ni et al., 2025).

Privacy protection. Although the source videos are publicly available, we treat privacy as a first-class concern. To mitigate re-identification risks in data release, we will: (1) publish a privacy-preserving version of the clips where human faces are blurred (including the interpreter and any other visible individuals); (2) avoid releasing personal identifiers beyond what is necessary for research (e.g., no names, no user handles), and provide only minimal source references (video ID and time range) for traceability; (3) explicitly prohibit any attempt to use LC-HKSLT for biometric identification, surveillance, or re-identification.

Copyright and licensing considerations. LC-HKSLT is derived from third-party online videos.

We do not claim ownership over the original content. Our release will follow a research-only license and will include clear attribution to the original sources, in line with broader concerns about unauthorized reuse of externally sourced data (Liu et al., 2025a). If any source content owner or appearing individual requests removal, we will promptly honor such requests.

Take-down policy. We will maintain a take-down mechanism: upon receiving a request with the source video identifier and timestamp range, we will remove the corresponding clips/annotations from future releases and provide an updated index to the community.

D Additional Quantitative Results

Fig. 3 visualizes the frame-level thought-to-time alignment of our latent reasoning module on five CSL-Daily examples. For each video, we show: (i) the attention heatmap from $K{=}8$ latent thoughts ($C_1{\sim}C_8$) to input frames, together with the selectivity entropy H and monotonic-violation rate `viol`; and (ii) the corresponding Chinese prediction with its English translation. The full model exhibits clear ordered bands (near-diagonal alignment), low `viol`, and relatively low H, indicating that different thoughts attend to progressively later evidence in a coherent manner. This observation is broadly consistent with the importance of spatially aligned and semantically discriminative representations in complex visual recognition settings (Yao et al., 2025; Zhang et al., 2025a). In contrast, removing $\mathcal{L}_{mono}$ disrupts temporal ordering (higher `viol`) and often leads to semantic reversals or role/polarity mistakes (e.g., predicting "can tell the difference" instead of "cannot"). Removing $\mathcal{L}_{cont}$ fragments attention across disjoint regions, correlating with disfluent or incomplete outputs. Replacing structured routing with soft routing produces diffuse, high-entropy allocations and more generic translations that miss key details. Overall, these cases qualitatively support that our chain-structured, regularized evidence allocation yields more faithful and grounded translations, especially for longer sentences requiring multi-step semantic composition. These visual patterns correspond closely to linguistic coherence in the generated translations: without the regularizers, models tend to swap event order or omit key arguments, confirming that chain-structured reasoning improves both alignment and semantic faithfulness (Jiang et al., 2026a).