Prompt, Pose, and Produce: Turkish Sign Language Generation with LLMs and Linguistically Motivated Pose Synthesis

Anonymous ACL submission

Abstract

Sign language generation for under-resourced languages remains challenging due to limited parallel corpora. We present a modular Turkish-to-Turkish Sign Language (TİD) generation framework with three central components: LLM-based glossing, spatial agreement rewriting, and context-aware out-of-vocabulary (OOV) resolution. The framework is built on TİD Sözlük, a community-developed dictionary of 9,082 isolated-sign and 2,346 sentence videos used for training and evaluation. We compare fine-tuned open LLMs (Gemma 3, Qwen 3) with prompted closed LLMs (Gemini 3 Flash, GPT-5.4) on 371 test sentences, finding that fine-tuned open models outperform prompted larger closed ones for Turkish-to-TİD glossing. Ablations show that spatial agreement and OOV resolution both significantly improve hand-trajectory alignment, with OOV resolution producing the larger effect. We release all code and data upon acceptance.

1 Introduction

Sign languages are natural languages with their own morphology, syntax, and spatial reference systems (Sandler and Lillo-Martin, 2006). Despite progress for American Sign Language (ASL) (Huenerfauth, 2008) and German Sign Language (DGS) (Saunders et al., 2020), Turkish Sign Language (TİD)[1] remains *under-represented* in computational research. Resources such as BosphorusSign (Camgöz et al., 2016), AUTSL (Sincan and Keles, 2020), and the community-developed TİD Sözlük (Makaroğlu and Dikyuva, 2017) exist, but computational technologies for TİD lag behind those for ASL, DGS, or British Sign Language (BSL). Prior TİD work has focused mainly on isolated sign recognition (Özdemir et al., 2020) and pose estimation (Keleş et al., 2023), with limited rule-based and statistical machine translation (Kayahan and Güngör, 2019). Furthermore, access to sign-language-mediated services remains limited: despite legislative changes, TİD has not been systematically incorporated into education or health-service provision, leaving Deaf signers underserved in key public settings (Kemaloğlu and Kemaloğlu, 2012). Sign language generation systems can help by translating written or spoken language into sign language video (Saunders et al., 2020; Walsh et al., 2024b), yet dominant methods have been developed mainly for higher-resource sign languages, leaving TİD-specific design choices underexplored. This gap matters because TİD is not signed Turkish, it has its own grammar and spatial mechanisms. It is also historically unrelated to the American and European sign languages that have been studied most extensively (Zeshan, 2002, 2003; Zwitserlood et al., 2012). Since TİD has limited parallel data and fully end-to-end neural generation can require substantial data and computation, we study a tractable gloss-to-pose route with explicit TİD-specific assumptions, beginning by evaluating recent LLMs for Turkish-to-TİD glossing. We organize our investigation around:

- **RQ1.** Can current LLMs produce TİD-conventional glosses for arbitrary Turkish input, and how do open and closed LLMs compare on this task?
- **RQ2.** Does explicit modeling of TİD-specific linguistic structure, especially spatial agreement and context-aware out-of-vocabulary (OOV) resolution, improve the quality of pose sequences over naive dictionary stitching?
- **RQ3.** For Turkish-to-TİD gloss translation, does parameter-level fine-tuning of open LLMs outperform in-context prompting of larger closed LLMs?

[1] TİD is the official language of the Deaf community of Türkiye, with an estimated signer population in the hundreds of thousands. Estimates place the deaf population in Türkiye at 187,500–337,500, with TİD signers corresponding to approximately 0.13%–0.27% of the total population (Kemaloğlu and Kemaloğlu, 2012).

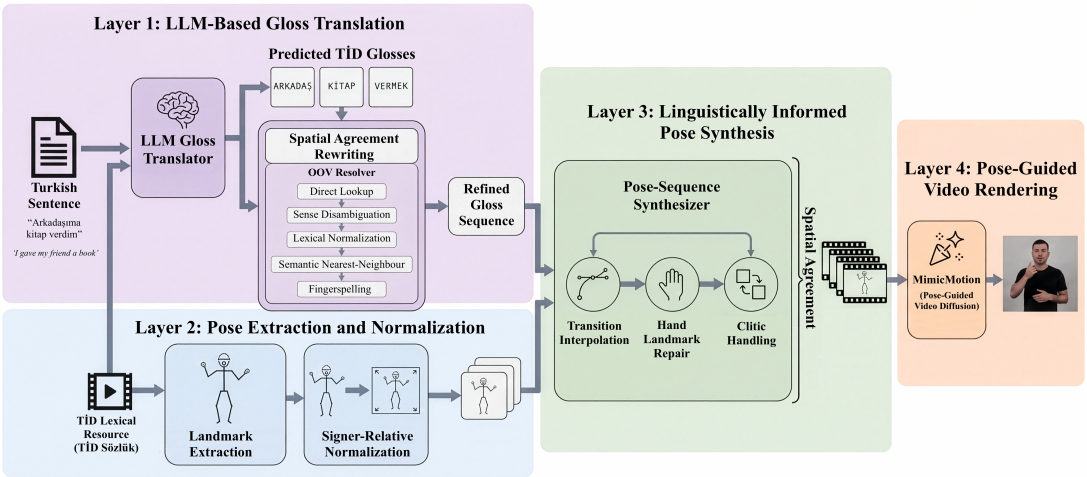

Figure 1: Architecture of the Turkish-text-to-TİD (Turkish Sign Language) video generation pipeline. Turkish input is converted into TİD glosses by fine-tuned LLMs. The predicted glosses are then passed through TİD-specific post-processing, including spatial agreement, OOV resolution, and dictionary lookup, before lexical sign clips are stitched into a pose sequence. The final stage uses the synthesized pose representation to guide video rendering.

To address these questions, we build on TİD Sözlük (Makaroğlu and Dikyuva, 2017), a dictionary resource developed with the Deaf community, and introduce three components: (i) fine-tuned and prompted LLMs for Turkish-to-gloss translation; (ii) deterministic post-processing for spatial agreement, context-aware OOV resolution, and dictionary-clip stitching; and (iii) pose-guided diffusion rendering (Figure 1). Quantitative evaluation focuses on glossing and pose synthesis: agreement inserts discourse loci and controls agreement movement and orientation, while OOV resolution makes context-sensitive lexical decisions with fingerspelling reserved as a final fallback. The contributions of this work are:

1. A *linguistically motivated spatial agreement* module that, drawing on the descriptive grammar of TİD (Kubuş, 2008; Sevinç, 2006), inserts discourse loci and rewrites the movement and orientation of verbs at the pose level.
2. A *context-aware OOV resolution* pipeline that resolves predicted glosses through direct lookup, sense disambiguation, lexical normalization, semantic nearest-neighbor substitution, and fingerspelling, yielding the largest improvement in hand-trajectory alignment.
3. A systematic study of *fine-tuning open LLMs* (Gemma 3 and Qwen 3 at multiple scales) against *in-context prompting of closed LLMs* (Gemini 3 Flash and GPT-5.4) for the low-resource Turkish-to-TİD gloss task, including ablations of the two modules above.

2 Related Work

Sign language generation sits at the intersection of low-resource translation, motion synthesis, and the grammatical description of individual sign languages. The work most directly relevant to this paper falls into two strands: pose- and video-based generation pipelines, and recent uses of LLMs for gloss-mediated translation. We summarize each and position our contributions with respect to them.

Sign Language Generation Early sign language generation systems relied on rule-based translation with avatar animation (Huenerfauth, 2008; Veale et al., 1998), often producing outputs that users perceived as rigid or unnatural. Statistical and neural methods later replaced these pipelines, including pose-based autoregressive generation (Saunders et al., 2020; Walsh et al., 2024a) and video synthesis via GANs or diffusion models (Fang et al., 2025; Qi et al., 2025). Several works model articulators beyond the hands, addressing non-manual markers (Zhang et al., 2025a), concept-motivated production for STEM (Sortino et al., 2026), and pragmatic articulation in classroom discourse (Imai et al., 2025b). More expressive body representations such as parametric 3D meshes have also been adopted (Pavlakos et al., 2019; Zuo et al., 2025). Upstream of these synthesis pipelines, self-supervised and contrastive representations have been developed for signs (Wong et al., 2025; Gueuwou et al., 2025b; Jiang et al., 2024; Gueuwou et al., 2025a), alongside linguistically grounded priors from phonology (Kezar et al., 2023, 2025b,a) and segmentation

(Moryossef et al., 2023). Evaluation for sign language generation is itself an active subarea, with recent work proposing semantically aware embedding metrics (Imai et al., 2025a) and dialogue-level production benchmarks for instruction generation (Inan et al., 2024). Dictionary-based production remains attractive for low-resource settings since it exploits existing lexical video, but naive concatenation produces discontinuities. Sign-stitching methods (Walsh et al., 2024b) normalize sign clips and interpolate between boundaries. Our system follows this dictionary-stitching line while adapting it to the lexical conventions and pose-level constraints required for TİD generation.

LLMs, Glossing, and Low-Resource Translation In-context machine translation with LLMs has recently been studied as a promising approach for low-resource MT, especially when prompts can incorporate dictionaries, grammar descriptions, or retrieved parallel examples (Tanzer et al., 2024b). A growing body of work applies them to sign language. Several efforts use LLMs as gloss-free or gloss-mediated translators (Wong et al., 2024; Gong et al., 2024; Chen et al., 2024; Fayyazsanavi et al., 2024), with Zhang et al. 2025a showing that GPT-4o supports English-to-ASL glossing under carefully designed prompting and Inan et al. (2025a) integrating multimodal sign processing directly into a pretrained LLM. Scaling studies have pushed sign language translation to larger corpora and broader domains (Zhang et al., 2024; Gueuwou et al., 2025a), while Tanzer et al. (2024a) revisit sentence-level evaluation assumptions that have shaped much of the SLT literature. Low-resource extensions cover African sign languages (Gueuwou et al., 2023), cross-corpus alignment for sign-to-sign translation (Inan et al., 2025b), dialogue-level instruction production (Inan et al., 2024), and bootstrapping annotations with sign language models for under-documented varieties (Lea et al., 2026). For TİD, we combine LLM-based gloss generation, comparing fine-tuned open LLMs against prompted closed LLMs, with deterministic post-processing rules derived from TİD Sözlük conventions and from the descriptive grammar of TİD.

3 Data and Preprocessing

This section describes the lexical and sentence-level resources our pipeline draws on and the preprocessing applied before the framework components in Section 4 operate.

3.1 Corpora

TİD Sözlük. We build on the publicly available TİD Sözlük resource[2] (Makaroğlu and Dikyuva, 2017), which offers 9,082 isolated-sign videos and 2,346 sentence videos accompanied by gloss annotations. This resource contains 2,000 lemmas selected based on frequency analysis from a corpus collected across 26 cities in Türkiye, involving 116 native TİD signers, making sure that broad geographic representativeness across regional signing variation.

İşaretçe. This open-source TİD resource[3] (which contains 2,411 signs) was used to augment and complement our TİD Sözlük data and to address out-of-vocabulary items through fingerspelling, the manual representation of written letters using hand configurations. İşaretçe alphabet clips are converted into wrist-relative canonical hand poses and then placed into the active signer frame during synthesis. Fingerspelling is used as the final fallback for unresolved lexical items after lexical and semantic substitution attempts.

3.2 TİD Gloss Conventions

We inspected TİD Sözlük sentence labels and use the following conventions for prompting and post-processing. TİD glosses are all-caps Turkish lemma/root forms; compounds are marked with ˆ, especially signs such as BILMEKˆDEĞIL (KNOWˆNOT)[4]; repetition and continuous aspect are represented by repeated glosses, and parenthesized tokens indicate optional material and are removed or separated during parsing. The corpus labels do not encode how signs are directed in the space, so the LLM is trained to produce corpus-style glosses; spatial marking is added later by the rewriting module described in Section 4.2.

3.3 Pose Extraction and Representation

We use MediaPipe Holistic (Lugaresi et al., 2019) to extract body and hand landmarks from TİD Sözlük videos, storing 33 pose landmarks and 21 landmarks per hand for each frame. We normalize pose coordinates relative to shoulder landmarks, apply temporal smoothing, validate pose completeness, and filter sequences with excessive missing skeletal points. The system then uses a signer-relative

[2] https://tidsozluk.aile.gov.tr/

[3] https://isaretce.com/

[4] We follow the sign-language linguistics convention of writing sign glosses in small caps to distinguish them from ordinary spoken or written words.

canonical frame to compensate for signer placement and camera angle variation. Each sign's temporal profile is trimmed using velocity and signing-space criteria to remove non-signing preparation and retraction.

3.4 Training, Validation, and Test Splits

The Turkish–TİD sentence pairs extracted from TİD Sözlük serve both as supervision for the gloss-translation models in Section 4 and as the basis for evaluation. We hold out 371 test sentences for evaluation. The remaining pairs are split into training (85%) and validation (15%) sets and used for all fine-tuning experiments; the same 371 test sentences are used for both gloss-level and pose-level evaluation.

4 Our Framework

Our framework treats Turkish-to-TİD generation as three coupled stages: gloss prediction from Turkish text, linguistically informed pose-sequence construction, and pose-guided video rendering. The first compares fine-tuned open LLMs with prompted closed LLMs; the second encodes TİD-specific structure through spatial agreement rewriting, context-aware OOV resolution, and clip stitching; the third renders the resulting pose sequence into video.

4.1 LLM-Based Gloss Translation

The first stage translates Turkish sentences into TİD glosses. We evaluate fine-tuned open models, Gemma 3 (Gemma Team et al., 2025) at 4B, 12B, and 27B and Qwen 3 (Yang et al., 2025) at 4B, 8B, and 32B, against Gemini 3 Flash (Google DeepMind, 2025) and GPT-5.4 (OpenAI, 2026) under one-shot and ten-shot prompting.

Open models are fine-tuned with bf16 precision, LoRA (rank 64, $\alpha = 16$, dropout 0.1) on attention and feed-forward projections, learning rate 2×10^{-4}, AdamW, warmup ratio 0.03, and 3 epochs; full hyperparameters are in Appendix C. Closed models are prompted only; system messages are reproduced in Appendix B.

4.2 Spatial Agreement Rewriting

We follow the standard distinction between plain, agreement, and spatial verbs (Padden, 1986), where plain verbs do not inflect for person or number. Agreement verbs can modify movement, hand orientation, or both so that the verb points from subject to object in regular agreement verbs and from object to subject in backwards agreement verbs. These loci may pick up physically present referents or abstract discourse points; TİD also licenses null arguments, whose distribution interacts with verb type (Sevinç, 2006).

The rewriting module assigns a signing-space locus to each argument: locus 1 is the signer, locus 2 is the addressee, and third-person referents receive loci on the left or right side of the signing space (3a, 3b, ...) introduced by pointing signs (IX_{3a}, IX_{3b}) after the gloss of the referring noun. When a verb from the agreeing-verb inventory is present and its arguments are identifiable, it is annotated in subscript notation (e.g., ${}_{3a}\text{GÖRMEK}_{3b}$); when both arguments are first- or second-person pronouns, the verb is directed between the signer and addressee loci directly without overt pointing signs (e.g., ${}_{2}\text{GÖRMEK}_{1}$). The subscript preceding the verb marks the subject locus, and the subscript following it marks the object locus. Examples (1)–(2) illustrate the rewriting for third-person referents and first/second-person pronouns:

(1) a. ADAM KADIN GÖRMEK
MAN WOMAN SEE
'The man saw the woman.'

b. ADAM IX_{3a} KADIN IX_{3b} ${}_{3a}\text{GÖRMEK}_{3b}$
MAN IX_{3a} WOMAN IX_{3b} ${}_{3a}\text{SEE}_{3b}$
'The man saw the woman.'

(2) a. SEN BEN GÖRMEK
YOU I SEE
'You saw me.'

b. ${}_{2}\text{GÖRMEK}_{1}$
${}_{2}\text{SEE}_{1}$
'You saw me.'

The agreeing-verb inventory of 60 verbs, compiled from (Kubuş, 2008) and listed in Appendix A, classifies each verb into one of three categories. In the pose domain, pointing signs are realized as index-finger extensions toward the assigned locus position without full-arm extension. For agreement verbs, the dictionary clip is first placed in a neutral canonical frame and citation-form rest phases are trimmed; the category-appropriate wrist trajectory and orientation adjustments are then applied to direct the movement between loci.

4.3 Context-Aware OOV Resolution

A central component of our framework is a context-aware out-of-vocabulary resolver that decides how each predicted gloss is realized in the pose domain. Because corpus-style glosses do not exhaustively

cover the TİD Sözlük inventory, many generated tokens are absent from the dictionary or polysemous within it. Naive lookup discards these tokens or returns an arbitrary sense, producing visible discontinuities. The resolver instead applies an ordered four-step strategy to every generated gloss token (summarized in Appendix D). Each successful match is routed through Step 1 so that polysemous matches still pass through Sense Disambiguation.

Step 1: Direct Lookup with Sense Disambiguation. The token is looked up verbatim in TİD Sözlük. If the entry is monosemous, its sign is returned directly. If the entry is polysemous, i.e., TİD Sözlük records more than one distinct sign for the same Turkish word, the resolver invokes **Sense Disambiguation**. For each candidate sign, we concatenate the dictionary description and example sentence, embed the result with *paraphrase-multilingual-mpnet-base-v2* (Reimers and Gurevych, 2019), and select the sense with highest cosine similarity to the full input Turkish sentence. This produces a per-occurrence sense decision conditioned on context rather than a corpus-wide default.

Step 2: Lexical Normalization. If Direct Lookup fails, the resolver applies deterministic rewrite rules: compound splitting on ˆ, joint-entry comma-alias matching, conservative spelling reduction, inflectional suffix stripping, and semantic-prefix lookup. Each successful rewrite re-enters Step 1, so polysemous normalized forms still benefit from Sense Disambiguation.

Step 3: Semantic Nearest-Neighbor. When no normalized form succeeds, the token together with its sentence context is embedded using the same *paraphrase-multilingual-mpnet-base-v2* model and compared with precomputed embeddings of dictionary glosses; the closest match is accepted only if cosine similarity ≥ 0.80. This conservative threshold avoids spurious substitutions: accepted tokens typically share a root or near-synonym relation with the retrieved entry (e.g., PARLAMAK 'to gleam' resolves to PARLAK 'bright/shiny'). The result also re-enters Step 1.

Step 4: Fingerspelling. Tokens with no dictionary coverage and no sufficiently close semantic neighbor are spelled letter-by-letter using the İşaretçe alphabet (Section 3). Some proper nouns, including personal names and brand names, as well as acronyms, are routed to this step.

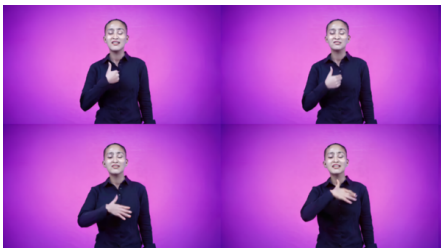

Figure 2: Samples from diffusion rendering stage.

Qualitative examples of each strategy are given in Appendix F.

4.4 Stitching and Synthesis

We next stitch the resolved sign clips into a continuous pose sequence. The system canonicalizes every clip into a signer-relative frame, rejects invalid pose frames, repairs collapsed hand frames, crops citation-form rest, and inserts transition frames between consecutive clips. The number of inserted transition frames is proportional to the *boundary wrist displacement*, the Euclidean gap between dominant-hand wrist positions at the clip boundary, so fast hand jumps receive longer interpolation and near-coincident endpoints receive only a brief transition. Inactive hands are filled with a stable neutral hand pose; torso and body landmarks are interpolated over time so body lines remain stable in the skeleton visualization.

Each sentence begins with arms released in a neutral position and transitions into the first gloss. Compounds such as BILMEKˆDEĞIL (KNOWˆNOT) are expanded into parts, but the following DEĞIL (NOT) is treated as a reduced clitic whose location assimilates to the previous sign rather than starting as a full citation form. Reduplicated signs are also handled by repeating the internal movement band.

4.5 Pose-Guided Rendering

Pose sequences are rendered into sign language video using MimicMotion (Zhang et al., 2025b), a video generation framework built on Stable Video Diffusion (Blattmann et al., 2023). We use MimicMotion as an off-the-shelf renderer: stitched pose sequences provide the motion condition, and a reference appearance frame taken from TİD Sözlük supplies a consistent signer identity. This stage demonstrates that the synthesized poses can drive video generation; quantitative video-quality evaluation and claims about photorealism are deferred to future work. Qualitative samples are in Figure 2.

5 Findings

We evaluate on the 371 test sentences in two settings: *gloss translation*, scoring LLM outputs against gold TİD glosses, and *end-to-end pose synthesis*, scoring stitched poses against ground-truth-gloss stitches and actual TİD Sözlük videos. We also ablate spatial agreement and OOV resolution.

5.1 Evaluation Metrics

Gloss translation. For RQ1 and RQ3, we score predicted glosses against TİD Sözlük references using **BLEU** (Papineni et al., 2002) at BLEU-1 through BLEU-4, **ROUGE** (Lin, 2004) at ROUGE-1, ROUGE-2, and ROUGE-L, and **semantic similarity**. Semantic similarity is mean cosine similarity between *paraphrase-multilingual-mpnet-base-v2* embeddings of predicted and reference glosses (Reimers and Gurevych, 2019).

Pose synthesis. For RQ2, we compare the pose sequences produced by the full pipeline against a target pose sequence using two metrics widely used in the sign language processing literature for evaluating pose generation quality. Because native-signer evaluation is outside the current study, we treat these metrics as proxy measures of pose alignment rather than comprehensibility. **Mean Per Joint Position Error (MPJPE)** measures the average Euclidean distance between joints over time:

$$\text{MPJPE} = \frac{1}{N \cdot T} \sum_{t=1}^{T} \sum_{j=1}^{N} \|p_j^t - \hat{p}_j^t\|_2, \quad (1)$$

where N is the number of joints, T the number of frames, and p_j^t and $\hat{p}_j^t$ the reference and predicted positions of joint j at frame t. **Dynamic Time Warping (DTW)** (Berndt and Clifford, 1994) measures temporal alignment quality:

$$\text{DTW}(X, Y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2}, \quad (2)$$

where X and Y are the predicted and reference pose sequences, π is the warping path, and $d(\cdot, \cdot)$ is Euclidean distance between pose vectors. We report both metrics over three landmark groups: **All** (all skeletal points), **Hands** (left- and right-hand keypoints only), and **Pose** (body landmarks only). For both metrics, lower is better.

5.2 Gloss Translation

Table 1 reports Turkish-to-TİD gloss translation across all ten LLM configurations; the dashed midrule separates fine-tuned open LLMs from prompted closed LLMs.

Fine-tuned open LLMs beat prompted closed LLMs. The strongest fine-tuned open model, Gemma 3 27B, reaches 17.31 BLEU-4 and 0.870 semantic similarity, compared with 14.54 BLEU-4 and 0.865 semantic similarity for the strongest prompted closed model, Gemini 3 Flash (10-shot). The fine-tuned mid-scale Gemma 3 12B (16.83 BLEU-4) also outperforms every prompted closed model. This pattern answers RQ3 in favor of parameter adaptation for the low-resource TİD gloss task, supporting our broader claim that TİD-specific structure is better learned through adaptation than supplied only through prompts.

Scaling within the Gemma family is consistent. Gemma 3 27B improves over 12B by 0.49 BLEU-4 and over 4B by 1.96 BLEU-4 and is the strongest model across all reported gloss metrics. Qwen 3 scales less monotonically (8B underperforms 4B on BLEU-4), and Qwen 3 32B never matches Gemma 3 27B. The contrast suggests that scale helps most when the underlying model family and fine-tuning setup can absorb the notation and ordering constraints of TİD glosses.

More in-context examples help, but do not close the gap. For both Gemini 3 Flash and GPT-5.4, moving from one-shot to ten-shot prompting improves BLEU-4 (by 0.82 and 0.77 respectively). The improvement is consistent but small relative to the gap to fine-tuned open models. Examples help with format, but do not replace task-specific adaptation for low-resource sign-language glossing.

5.3 End-to-End Pose Synthesis

We evaluate the full pipeline on the same 371 test sentences using Gemma 3 27B, Qwen 3 32B, GPT-5.4, and Gemini 3 Flash. For closed models, we use the stronger 10-shot condition. We first compare model-predicted-gloss poses against ground-truth-gloss poses passed through the same stitching pipeline, isolating gloss prediction errors. We then compare stitched outputs against actual TİD Sözlük sentence videos, a stricter setting that includes the gap between dictionary stitching and natural sentence signing. Figure 3 summarizes hand-trajectory metrics; full MPJPE and DTW appear in Appendix E.

Predicted-gloss stitches closely approximate ground-truth-gloss stitches. All four systems

Model	Setting	B-1	B-2	B-3	B-4	R-1	R-2	R-L	Sem.
Gemma 3 4B	fine-tuned	46.57	30.30	20.94	$15.35_{\pm 1.4}$	0.546	0.250	0.492	0.839
Gemma 3 12B	fine-tuned	51.10	33.71	23.08	$16.83_{\pm 1.5}$	0.594	0.279	0.532	0.861
Gemma 3 27B	fine-tuned	**51.26**	**34.01**	**23.55**	$\mathbf{17.31}_{\pm 1.5}$	**0.603**	**0.284**	**0.543**	**0.870**
Qwen 3 4B	fine-tuned	45.70	29.49	20.13	$14.72_{\pm 1.4}$	0.546	0.242	0.491	0.837
Qwen 3 8B	fine-tuned	45.59	28.73	19.36	$13.98_{\pm 1.2}$	0.544	0.234	0.487	0.842
Qwen 3 32B	fine-tuned	49.76	32.42	22.29	$16.44_{\pm 1.5}$	0.585	0.265	0.529	0.854
Gemini 3 Flash	1-shot	49.79	29.37	19.23	$13.72_{\pm 1.2}$	0.581	0.222	0.506	0.865
Gemini 3 Flash	10-shot	50.99	30.63	20.29	$14.54_{\pm 1.2}$	0.585	0.231	0.514	0.865
GPT-5.4	1-shot	48.42	28.07	18.29	$13.09_{\pm 1.3}$	0.565	0.210	0.492	0.861
GPT-5.4	10-shot	50.14	29.61	19.43	$13.86_{\pm 1.3}$	0.574	0.221	0.504	0.866

Table 1: Turkish-to-TİD gloss translation on the 371-sentence test set. Top block: fine-tuned open LLMs (LoRA). Bottom block (below dashed line): prompted closed LLMs at one-shot and ten-shot. BLEU-4 subscripts show 95% bootstrap CI half-widths. For all metrics, higher is better; best value per column is in bold.

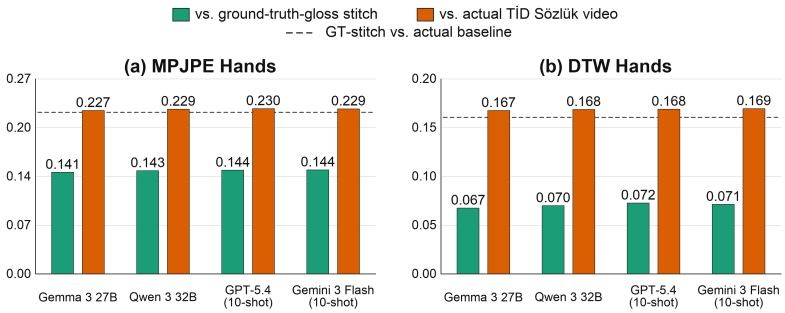

Figure 3: Pose metrics (lower is better). Teal: predicted-gloss stitch vs. ground-truth-gloss stitch. Orange: full pipeline vs. actual TİD Sözlük video. Dashed line: ground-truth-gloss stitch vs. actual video.

produce predicted-gloss pose sequences close to the ground-truth-gloss synthesis (Figure 3, teal bars; Appendix E), with Gemma 3 27B best across all six metrics. System differences are small in absolute terms (MPJPE-All within 0.0036), indicating that gloss prediction appears to be a smaller pose-error source than synthesis mismatch among the strongest evaluated systems. In terms of the pipeline argument, this shifts the main bottleneck from LLM glossing toward how dictionary signs are synthesized in sentence context.

A substantial gap remains between dictionary stitching and natural signing. Against actual TİD Sözlük videos (Figure 3, vermillion bars; Appendix E), Gemma 3 27B again leads among model outputs, but even the ground-truth gloss stitch (DTW Hands ≈ 0.16, MPJPE Hands ≈ 0.22) sits well above zero. This gap reflects timing, coarticulation, prosody, and signer-specific realization absent from isolated lexical clips. The result therefore qualifies the promise of modular generation: lexical resources make TİD synthesis feasible, but naturalness depends on modeling sentence-level signing beyond dictionary lookup.

5.4 Ablations: Agreement and OOV

To isolate the two linguistically motivated components (RQ2), we run ablations against the full pipeline, focusing on **DTW Hands** in the actual-video comparison. Agreement is evaluated on the 169 test sentences whose reference gloss contains an agreeing verb.

Spatial agreement improves hand-trajectory alignment. Disabling agreement produces small but consistent DTW Hands degradations of 0.0003–0.0018 across model outputs on the affected subset. The effect is significant after Holm correction for Gemma 3 27B ($p_{\text{Holm}} = .0063$), GPT-5.4 ($p_{\text{Holm}} = .0060$), and Gemini 3 Flash ($p_{\text{Holm}} = .0003$), while Qwen 3 32B is smaller and non-significant ($p_{\text{Holm}} = .1102$). The ground-truth gloss stitch shows the same direction (ΔDTW Hands = 0.0019; $p_{\text{Holm}} < .0001$). This supports RQ2 by showing that even a compact, deterministic treatment of TİD agreement contributes measurable structure beyond naive clip stitching.

Context-aware OOV resolution is the larger contributor. OOV resolution activates for 10.6% of Gemma 3 27B tokens and 10.2% of Qwen 3 32B

tokens, with another 7.5% in-vocabulary but polysemous (Appendix D). Semantic Nearest-Neighbor accounts for the largest OOV share (≈44%), followed by Lexical Normalization (≈31%) and Fingerspelling (≈26%). On sentences where the full pipeline used Sense Disambiguation or Semantic Nearest-Neighbor (261–281 samples), disabling the contextual resolver worsens DTW Hands by 0.0038–0.0047, with significant effects for all four systems after Holm correction ($p_{\text{Holm}} \leq 3.5\times10^{-7}$). The ground-truth gloss stitch shows the same pattern (ΔDTW Hands = 0.0051; $p_{\text{Holm}} < .0001$), while MPJPE Hands again misses this trajectory-level degradation. As the examples in Appendix F show, OOV resolution improves coverage but cannot repair upstream glossing errors. In sum, the ablations show that lightweight linguistic modules recover useful TİD structure, but still depend on gloss quality.

6 Discussion

RQ1, RQ3: Fine-tuning beats in-context prompting for low-resource sign-language glossing. Gemma 3 27B achieves 17.31 BLEU-4 and 0.833 semantic similarity, compared with 14.54 BLEU-4 and 0.826 for the strongest prompted closed model, Gemini 3 Flash. Even Gemma 3 12B (16.83 BLEU-4) outperforms every prompted closed configuration, consistent with MT work showing that parameter-efficient fine-tuning can outperform zero- and few-shot prompting (Zhang et al., 2023). This also aligns with recent sign-language gloss-translation work that adapts pre-trained LLMs through fine-tuning, although prior work has mostly focused on gloss-to-text rather than text-to-gloss translation (Fayyazsanavi et al., 2024).We attribute this to TİD-specific gloss conventions: visual-gestural ordering, all-caps lemmas, compound markers, and omitted inflection. Fine-tuning appears to learn these notation and ordering patterns more reliably than prompting.

RQ2: Linguistic structure earns its place in the pipeline. Both spatial agreement and context-aware OOV resolution improve hand-trajectory alignment where they apply, with OOV resolution producing the larger effect (DTW Hands improvement of 0.0038–0.0047 vs. a small but consistent 0.0003–0.0018 for agreement). MPJPE Hands is largely insensitive because these modules reshape *which* trajectory the hand follows rather than frame-wise joint distance, while DTW Hands captures the trajectory-level effect.

Pipeline bottleneck. The pose comparisons locate the pipeline bottleneck. All four systems land within MPJPE-All ≈ 0.0036 of each other in the intra-pipeline comparison, so gloss prediction appears to be a smaller pose-error source than synthesis mismatch among the strongest evaluated systems. The actual-video comparison gives higher errors even for the ground-truth gloss stitch baseline (DTW Hands ≈ 0.16), showing that timing, coarticulation, prosody, and signer-specific realization cannot be recovered from isolated lexical signs and corpus-style glosses. This is consistent with recent ASL generation work emphasizing that understandable sign-language video requires not only manual gloss realization but also non-manual markers, facial expressions, and motion naturalness (Zhang et al., 2025a). This distinction matters for future work: improving gloss accuracy alone will yield diminishing pose gains unless the synthesis stage also learns how signs deform, shorten, and coordinate in sentence context.

Implications for low-resource sign languages. The modular architecture is practical for low-resource settings because it can reuse a dictionary corpus, compact linguistic rules, and modern LLM glossers. Its main advantage is controllability: each failure can be traced to glossing, lexical lookup, agreement placement, stitching, or rendering rather than hidden inside a single end-to-end model. Yet dictionary stitching is not natural signing; future gains will require better modeling of coarticulation, non-manual markers, and signer-specific timing. Our pipeline can also be extended to other under-resourced sign languages.

7 Conclusion

We present, to our knowledge, the first modular Turkish-text-to-TİD pipeline combining LLM-based glossing, TİD-specific pose synthesis, and pose-guided rendering. Gemma 3 27B performs best for gloss translation, and ablations show that spatial agreement and context-aware OOV resolution improve hand-trajectory alignment over naive dictionary stitching. More broadly, the results show that dictionary resources, compact linguistic rules, and adaptable LLMs can be practical for under-resourced sign-language generation.

8 Limitations

This study has several limitations. First, and most critically, no human evaluation is included. MPJPE and DTW isolate geometric and temporal differences but do not capture whether a generated TİD sentence is understandable, natural, culturally appropriate, or acceptable to native signers. The photorealistic rendering component is demonstrated qualitatively only; video-quality metrics (e.g., temporal coherence, perceptual realism) and native-signer intelligibility judgments on rendered video are deferred to future work. Human evaluation with Deaf TİD users is necessary before conclusions about real-world usability can be drawn.

Second, the pose-stitching pipeline depends on TİD Sözlük and related lexical resources. This makes the system practical in a low-resource setting, but constrains the output to dictionary-style lexical signs. The actual-video comparison shows that even ground-truth gloss stitches differ from natural sentence videos, reflecting missing coarticulation, prosody, non-manual markers, discourse structure, and signer-specific timing.

Third, the spatial agreement and OOV modules are deterministic approximations. They do not model the full linguistic range of TİD agreement, classifier constructions, role shift, or context-dependent reference tracking, and OOV substitution may be imperfect in specific discourse contexts. Future work could also add an iconicity-aware OOV strategy. Rather than resolving every missing lexical item through textual nearest-neighbor substitution or fingerspelling, the resolver could compose provisional visually motivated forms when appropriate, using iconic mappings between meaning and articulatory form, such as arm movement for TO-FLY or a claw/round handshape for classifying an object like BALL (Keleş et al., 2026).

Finally, the model comparison is limited to the available 371-sentence test set and selected LLM configurations; results may change with larger corpora, stronger renderers, or evaluation sets targeting phenomena such as classifier predicates, negation, fingerspelling, and regional variation.

References

Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pages 359–370.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *Preprint*, arXiv:2311.15127.

Necati Cihan Camgöz, Ahmet Alp Kındıroğlu, Serpil Karabüklü, Meltem Kelepir, Ayşe Sumru Özsoy, and Lale Akarun. 2016. BosphorusSign: A Turkish Sign Language recognition corpus in health and finance domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1383–1388, Portorož, Slovenia. European Language Resources Association (ELRA).

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081, Torino, Italia.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. 2025. Signdiff: Diffusion model for american sign language production. *Preprint*, arXiv:2308.16082.

Pooya Fayyazsanavi, Antonios Anastasopoulos, and Jana Kosecka. 2024. Gloss2Text: Sign language gloss translation using LLMs and semantically aware label smoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16162–16171, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are Good Sign Language Translators . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372, Los Alamitos, CA, USA. IEEE Computer Society.

Google DeepMind. 2025. Gemini 3 flash – model card. `https://deepmind.google/models/gemini/flash/`.

Shester Gueuwou, Xiaodan Du, Gregory Shakhnarovich, and Karen Livescu. 2025a. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521, Vienna, Austria. Association for Computational Linguistics.

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

Shester Gueuwou, Kate Takyi, Mathias Müller, Marco Stanley Nyarko, Richard Adade, and Rose-Mary Owusuaa Mensah Gyening. 2023. AfriSign: Machine translation for African sign languages. In *Proceedings of the 4th Workshop on African Natural Language Processing (AfricaNLP)*.

Matt Huenerfauth. 2008. Generating American Sign Language animation: Overcoming misconceptions and technical challenges. *Universal Access in the Information Society*, 6(4):419–434.

Saki Imai, Mert Inan, Anthony B. Sicilia, and Malihe Alikhani. 2025a. SiLVERScore: Semantically-aware embeddings for sign language generation evaluation. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing — Natural Language Processing in the Generative AI Era*, pages 452–461, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Saki Imai, Lee Kezar, Laurel Aichler, Mert Inan, Erin Walker, Alicia Wooten, Lorna Quandt, and Malihe Alikhani. 2025b. How pragmatics shape articulation: A computational case study in STEM ASL discourse. arXiv preprint arXiv:2510.23842.

Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt, and Malihe Alikhani. 2024. Generating signed language instructions in large-scale dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 140–154, Mexico City, Mexico. Association for Computational Linguistics.

Mert Inan, Anthony Sicilia, and Malihe Alikhani. 2025a. SignAlignLM: Integrating multimodal sign language processing into large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3691–3706, Vienna, Austria. Association for Computational Linguistics.

Mert Inan, Yang Zhong, Vidya Ganesh, and Malihe Alikhani. 2025b. How to align multiple signed language corpora for better sign-to-sign translations? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4003–4016, Albuquerque, New Mexico. Association for Computational Linguistics.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, Miami, Florida, USA. Association for Computational Linguistics.

Dilek Kayahan and Tunga Güngör. 2019. A hybrid translation system from Turkish spoken language to Turkish Sign Language. In *2019 IEEE international symposium on innovations in intelligent systems and applications (inista)*, pages 1–6. IEEE.

Onur Keleş, Ceren Oksal, and Emre Bilgili. 2023. Using pose estimation for reference tracking in Turkish Sign Language. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Onur Keleş, Aslı Özyürek, Gerardo Ortega, Kadir Gökgöz, and Esam Ghaleb. 2026. The visual iconicity challenge: Evaluating vision-language models on sign language form-meaning mapping. *Preprint*, arXiv:2510.08482.

Yusuf Kemal Kemaloğlu and Pınar Yaprak Kemaloğlu. 2012. The history of sign language and deaf education in Turkey. *The Turkish Journal of Ear Nose and Throat*, 22(2):65–76.

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

Lee Kezar, Zed Sevcikova Sehyr, and Jesse Thomason. 2025b. Phonological representation learning for isolated signs improves out-of-vocabulary generalization. arXiv preprint arXiv:2509.04745.

Lee Kezar, Jesse Thomason, and Zed Sevcikova Sehyr. 2023. Improving sign recognition with phonology. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737, Dubrovnik, Croatia. Association for Computational Linguistics.

Okan Kubuş. 2008. An analysis of Turkish Sign Language (TİD) phonology and morphology. Master's thesis, Middle East Technical University, Ankara, Turkey.

Colin Lea, Vasileios Baltatzis, Connor Gillis, Raja Kushalnagar, Lorna Quandt, and Leah Findlater. 2026. Bootstrapping sign language annotations with sign language models. arXiv preprint arXiv:2604.07606 (accepted to CVPR 2026).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *Preprint*, arXiv:1906.08172.

Bahtiyar Makaroğlu and Hasan Dikyuva, editors. 2017. *Güncel Türk İşaret Dili Sözlüğü*. Aile ve Sosyal Hizmetler Bakanlığı, Ankara, Turkey.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

OpenAI. 2026. Gpt-5.4 thinking system card. `https://openai.com/index/gpt-5-4-thinking-system-card`.

Oğulcan Özdemir, Ahmet Alp Kındıroğlu, Necati Cihan Camgöz, and Lale Akarun. 2020. BosphorusSign22k sign language recognition dataset. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 181–188, Marseille, France. European Language Resources Association (ELRA).

Carol Padden. 1986. Verbs and role-shifting in american sign language. In *Proceedings of the fourth national symposium on sign language research and teaching*, volume 44, page 57. National Association of the Deaf Silver Spring, MD.

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

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. *Preprint*, arXiv:1904.05866.

Fan Qi, Yu Duan, Huaiwen Zhang, and Changsheng Xu. 2025. Signgen: End-to-end sign language video generation with latent diffusion. In *Computer Vision – ECCV 2024*, pages 252–270, Cham. Springer Nature Switzerland.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. *Preprint*, arXiv:2004.14874.

Ayça Müge Sevinç. 2006. Grammatical relations and word order in Turkish Sign Language (TİD). Master's thesis, Middle East Technical University, Ankara, Turkey.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. AUTSL: a large scale multi-modal Turkish Sign Language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Rachel Sortino, Christina Kim, Thalia Guettler, Katie McClyman, Lorna Quandt, Alicia Wooten, and Rachel Pizzie. 2026. Producing concept-motivated signs supports learning of STEM in American Sign Language. *npj Science of Learning*.

Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. 2024a. Reconsidering sentence-level sign language translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6262–6287, Miami, Florida, USA. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024b. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

Tony Veale, Alan Conway, and Bróna Collins. 1998. The challenges of cross-modal translation: English-to-sign-language translation in the Zardoz system. *Machine Translation*, 13:81–106.

Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, and Richard Bowden. 2024a. A data-driven representation for sign language production. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10.

Harry Walsh, Ben Saunders, and Richard Bowden. 2024b. Sign stitching: A novel approach to sign language production. *Preprint*, arXiv:2405.07663.

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen 3 technical report. *Preprint*, arXiv:2505.09388.

Ulrike Zeshan. 2002. Sign language in Turkey: The story of a hidden language. *Turkic Languages*, 6:229–274.

Ulrike Zeshan. 2003. Aspects of Türk İşaret Dili (Turkish Sign Language). *Sign Language & Linguistics*, 6(1):43–75.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. *Preprint*, arXiv:2407.11855.

Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna C Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. 2025a. Towards AI-driven sign language generation with non-manual markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. 2025b. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *Preprint*, arXiv:2406.19680.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. 2025. Signs as tokens: A retrieval-enhanced multilingual sign language generator. *Preprint*, arXiv:2411.17799.

Inge Zwitserlood, Pamela Perniss, and Aslı Özyürek. 2012. An empirical investigation of expression of multiple entities in Turkish Sign Language (TİD): Considering the effects of modality. *Lingua*, 122(14):1636–1667.

A TİD Agreement Verb Inventory

Table 2 lists the verbs in the TİD agreement inventory.

Category 1: movement and orientation (40 verbs)					
ALAY ETMEK	*mock*	ALMAK	*take*	AVANTAJ SAĞLAMAK	*benefit (s.o.)*
AYRILMAK	*separate*	BAĞIRMAK	*shout at*	BAKMAK	*look at*
BAŞVURMAK	*apply to*	BIÇAKLAMAK	*stab*	ÇAĞIRMAK	*call/summon*
DESTEKLEMEK	*support*	DÖVMEK	*beat*	DURDURMAK	*stop*
E-POSTA ATMAK	*email*	EMRETMEK	*order*	ETKILEMEK	*affect*
FAKS GÖNDERMEK	*fax*	GÖNDERMEK	*send*	GÖRMEK	*see*
IĞNELEMEK	*tease/sting*	ISIRMAK	*bite*	KARŞI OLMAK	*oppose*
KOPMAK	*break off*	KOPYA ÇEKMEK	*copy/cheat*	KOVMAK	*dismiss*
KÖTÜLEMEK	*slander*	KÜFÜR ETMEK	*curse at*	ÖDEMEK	*pay*
ÖĞRETMEK	*teach*	ÖLDÜRMEK	*kill*	ÖPMEK	*kiss*
RICA ETMEK	*request*	SEÇMEK	*choose*	SESLENMEK	*call out to*
SUSTURMAK	*silence*	ŞIKAYET ETMEK	*complain*	TAKIP ETMEK	*follow*
TEKLIF ETMEK	*offer*	VURMAK	*hit*	YENMEK	*defeat*
ZORLAMAK	*force*				
Category 2: orientation only (3 verbs)					
HOŞLANMAK	*like*	PAYLAŞMAK	*share*	SORGULAMAK	*question*
Category 3: trajectory only (17 verbs)					
ANLATMAK	*explain*	ARAMAK	*call/search*	BESLEMEK	*feed*
BORÇ VERMEK	*lend*	CEVAPLAMAK	*answer*	DAVET ETMEK	*invite*
HABER VERMEK	*notify*	KUTLAMAK	*congratulate*	ÖDÜL VERMEK	*award*
SATMAK	*sell*	SÖYLEMEK	*tell*	SORU SORMAK	*ask*
SUÇ ATMAK	*accuse*	TEŞEKKÜR ETMEK	*thank*	VERMEK	*give*
YÖNETMEK	*manage*	ZULÜM ETMEK	*oppress*		

Table 2: TİD agreement verb inventory (60 verbs) taken from (Kubuş, 2008). Category 1 modifies both wrist trajectory and wrist-yaw orientation; Category 2 modifies orientation only; Category 3 modifies trajectory only.

B Full System Message

```
You are a model that converts Turkish
↪  sentences into Turkish Sign Language
↪  (TİD) gloss sequences.

Task:
Given one Turkish sentence, output
↪  exactly one line containing the
↪  corresponding TİD glosses.

Output constraints:
- Output only the gloss sequence.
- Separate glosses with single spaces.
- Do not add explanations, punctuation,
↪  quotes, or extra lines.

Gloss conventions:
- Use ALL-CAPS Turkish lemma/root
↪  glosses.
- Do not output lowercase forms or
↪  Turkish inflectional suffixes.
- Avoid Turkish participial or suffixed
↪  surface forms as one-off gloss
↪  tokens when the line should use
↪  standard citation lemmas instead.
- Prefer TİD-style information
↪  structure:
  - time/topic first when present
  - then Subject-Object-Verb when the
  ↪  clause has agent and patient
- Omit tense, agreement, person, and
↪  case morphology when they are not
↪  lexical signs.
- Use `^` only for fixed lexical
↪  compounds or lexicalized negation.
↪  For lexicalized negation the usual
↪  form is one compound token: the
↪  negated lemma, then `^`, then DEĞİL
↪  (i.e. LEMMA^DEĞİL).
- Do not output optional alternatives or
↪  parenthesized material.
- If a source expression is idiomatic,
↪  output the most natural TİD gloss
↪  sequence rather than a word-for-word
↪  mapping.
```

C Fine-tuning Hyperparameters

Table 3 lists the hyperparameters shared across all fine-tuned open models. Target modules are the attention and feed-forward projection matrices.

Hyperparameter	Value
Precision	bf16
LoRA rank	64
LoRA α	16
LoRA dropout	0.1
Learning rate	2×10^{-4}
Optimizer	AdamW
Warmup ratio	0.03
Training epochs	3
Max sequence length	1024

Table 3: Fine-tuning hyperparameters shared across all open models.

D OOV Handling Statistics

Tables 4 and 5 report OOV resolution statistics. Table 4 shows mean strategy counts across Gemma 3 27B and Qwen 3 32B; Table 5 shows per-model OOV rates for all four systems. Rows above the first rule in Table 4 are in-vocabulary resolutions; rows below require OOV handling.

Resolution strategy	Mean Count	% of OOV	% of Total
Direct Lookup	4128	–	82.1%
+ Sense Disambiguation	377	–	7.5%
Lexical Normalisation	160	30.6%	3.2%
Semantic Nearest-Neighbour	228	43.6%	4.5%
Fingerspelling	135	25.8%	2.7%
Total Generated Tokens	5028	–	100.0%

Table 4: Mean resolution strategy distribution across Gemma 3 27B and Qwen 3 32B (≈5,028 tokens per system). Direct Lookup includes tokens found verbatim in TİD Sözlük; the indented Sense Disambiguation sub-row counts those additionally refined by polysemy resolution.

Model	Total Tokens	OOV Count	% OOV
Gemma 3 27B	4,965	525	10.6%
Qwen 3 32B	5,091	519	10.2%
GPT-5.4	5,536	666	12.0%
Gemini 3 Flash	5,190	562	10.8%

Table 5: Per-model OOV rates across the 371 test sentences for all four pipeline systems.

E Detailed Pose Evaluation Results

This appendix reports the full per-landmark MPJPE and DTW behind Figure 3. Table 6 compares pose sequences generated from each system's predicted glosses against pose sequences generated from the ground-truth glosses, with both sides passed through the same pose-stitching pipeline. Table 7 compares the full pipeline outputs against actual TİD Sözlük sentence videos. MPJPE and DTW are reported for all skeletal points (All), hands (H), and body landmarks (P). For all metrics, lower is better; the best value per column is in bold.

Model	**MPJPE-All**	**MPJPE-H**	**MPJPE-P**	**DTW-All**	**DTW-H**	**DTW-P**
Gemma 3 27B	**0.1125**	**0.1412**	**0.0644**	**0.0564**	**0.0673**	**0.0369**
Qwen 3 32B	0.1146	0.1433	0.0663	0.0586	0.0696	0.0391
GPT-5.4	0.1159	0.1440	0.0686	0.0616	0.0724	0.0425
Gemini 3 Flash	0.1161	0.1445	0.0683	0.0603	0.0710	0.0415

Table 6: Pose evaluation against the *ground-truth-gloss stitch*, i.e., the pose sequence obtained by stitching the reference TİD Sözlük glosses through the same pipeline. This comparison isolates the effect of gloss prediction errors under a shared synthesis pipeline.

Model	**MPJPE-All**	**MPJPE-H**	**MPJPE-P**	**DTW-All**	**DTW-H**	**DTW-P**
GT gloss stitch	**0.1799**	**0.2245**	**0.1049**	**0.1344**	**0.1599**	**0.0873**
Gemma 3 27B	0.1817	0.2269	0.1056	0.1405	0.1669	0.0900
Qwen 3 32B	0.1833	0.2287	0.1071	0.1409	0.1679	0.0908
GPT-5.4	0.1848	0.2297	0.1094	0.1422	0.1684	0.0932
Gemini 3 Flash	0.1843	0.2293	0.1087	0.1422	0.1689	0.0925

Table 7: Pose evaluation against *actual TİD Sözlük sentence videos*. The first row reports the ground-truth-gloss stitch's own distance to the actual video, i.e., the irreducible floor for any dictionary-style stitching pipeline; subsequent rows report each LLM-driven pipeline.

F Model Output Examples and Error Analysis

Table 8 reports four test-set sentences in which the glosses predicted by Gemma 3 27B require non-trivial OOV resolution before synthesis. The purpose of these examples is not to evaluate Gemma's overall translation quality, but to show how the downstream resolver handles predicted tokens that are missing from the TİD Sözlük inventory. Each predicted token was passed through the same resolver used by the synthesizer (Sec. D); the annotations in the rightmost column report the strategy selected for the relevant OOV token and the dictionary entry to which it was resolved.

Strategies in Examples. In Ex. 1 and Ex. 2 Gemma 3 27B predicts ARAMAK (to call) in both cases but the context differs; Sense Disambiguation selects a different sign variant for each (search vs. telephone call). Ex. 3 illustrates Semantic Nearest-Neighbour: PARLAMAK (to gleam) is absent from TİD Sözlük but is resolved to the adjective PARLAK (bright/shiny), which shares the same lexical root (cosine similarity 0.95). Ex. 4 shows Fingerspelling applied to the city name MARDIN, which has no dictionary sign.

Ex. 1	TR: *Ceket ve gömlek aldıktan sonra çizgili kravat aradım. Biraz aradıktan sonra bulunca iki tane aldım.* EN: *After buying a jacket and shirt I looked for a striped tie. After searching a while I found one and bought two.*
Reference	CEKET GÖMLEK BEN ALMAK KRAVAT ÇİZGİLİ BAKMAK ARAMAK ARAMAK BULMAK İKİ TANE ALMAK GL: JACKET SHIRT 1SG BUY TIE STRIPED LOOK SEARCH SEARCH FIND TWO TAKE
Gemma 3 27B	CEKET GÖMLEK ALMAK SONRA KRAVAT ÇİZGİLİ ARAMAK ARAMAK BULMAK İKİ TANE ALMAK GL: JACKET SHIRT BUY THEN TIE STRIPED SEARCH SEARCH FIND TWO TAKE
OOV	ARAMAK is polysemous (4 senses in TİD Sözlük); Sense Disambiguation selects the *search for something* variant.
Ex. 2	TR: *Sokakta yürürken yaşlı bir adamın düştüğünü gördüm. Hemen ambulansı aradım ve beş dakikada geldiler.* EN: *While walking on the street I saw an elderly man fall. I immediately called an ambulance and they arrived in five minutes.*
Reference	BEN SOKAK YÜRÜMEK YÜRÜMEK YAŞLI YÜRÜMEK DÜŞMEK BEN HEMEN AMBULANS ARAMAK BEŞ DAKİKA GELMEK GL: 1SG STREET WALK WALK ELDERLY WALK FALL 1SG IMMEDIATELY AMBULANCE CALL FIVE MINUTE ARRIVE
Gemma 3 27B	BEN SOKAK BEN YÜRÜMEK YÜRÜMEK YAŞLI ADAM DÜŞMEK BEN AMBULANS ARAMAK BEŞ DAKİKA GELMEK GL: 1SG STREET 1SG WALK WALK ELDERLY MAN FALL 1SG AMBULANCE CALL FIVE MINUTE ARRIVE
OOV	ARAMAK is polysemous; Sense Disambiguation selects the *call someone by phone* variant.
Ex. 3	TR: *Evimin duvarları çok kirlendiği için yeni boya aldım ve boyadım. Şimdi çok temiz oldu, her yer parlıyor.* EN: *My house walls had gotten very dirty so I bought new paint and painted them. Now it is very clean and everything is gleaming.*
Reference	BEN EV DUVAR HEPSİ KİRLENMEK BEN EN İYİ YENİ BOYA ALMAK BOYAMAK BOYAMAK TEMİZ PARLAMAK GL: 1SG HOUSE WALL ALL DIRTY 1SG BEST NEW PAINT BUY PAINT PAINT CLEAN GLEAM
Gemma 3 27B	BEN EV DUVAR KİR BEN BOYA ALMAK YENİ BOYA TEMİZLEMEK TEMİZLEMEK PARLAMAK GL: 1SG HOUSE WALL DIRT 1SG PAINT BUY NEW PAINT CLEAN CLEAN GLEAM
OOV	PARLAMAK → Semantic Nearest-Neighbour → PARLAK (cosine sim. 0.95; verb form resolved to dictionary adjective of same root).
After OOV	BEN EV DUVAR KİR BEN BOYA ALMAK YENİ BOYA TEMİZLEMEK TEMİZLEMEK PARLAK GL: 1SG HOUSE WALL DIRT 1SG PAINT BUY NEW PAINT CLEAN CLEAN BRIGHT
Ex. 4	TR: *Otuz yıldır Mardin'de oturuyorum.* EN: *I have been living in Mardin for thirty years.*
Reference	BEN MARDİN OTUZ SENE OTURMAK GL: 1SG MARDIN THIRTY YEAR LIVE
Gemma 3 27B	BEN MARDİN YER OTUZ SENE OTURMAK GL: 1SG MARDIN PLACE THIRTY YEAR LIVE
OOV	MARDIN → Fingerspelling (city name; no dictionary sign).
After OOV	BEN [FS:MARDİN] YER OTUZ SENE OTURMAK GL: 1SG [M-A-R-D-I-N] PLACE THIRTY YEAR LIVE

Table 8: OOV resolution examples from the test set (Gemma 3 27B predictions). *Sense Disambiguation* selects the contextually appropriate sign variant for polysemous entries (Ex. 1–2: same gloss token ARAMAK, different sign video). *Semantic Nearest-Neighbour* resolves absent tokens to the nearest dictionary entry by embedding similarity (Ex. 3: PARLAMAK→PARLAK). *Fingerspelling* uses the İşaretçe alphabet for proper nouns and final-fallback items (Ex. 4). GL lines provide English gloss translations. Underlined tokens indicate OOV or polysemy resolution.