

WSLP-Task2 at WSLP@EMNLP 2026: Gated Vision–Language Adjudication and Multi-Expert Fusion for Isolated Sign Language Recognition

Bolisetty Samanvitha and Ojas Srivastava

Pruthwik Mishra and Praveen Kumar Chandaliya

Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India

{u24ai024, u24ai028, pruthwikmishra, pkc}@aid.svnit.ac.in

Abstract

We present WSLP-Task2 for the WSLP@EMNLP 2026 isolated sign language recognition shared task. Given a short Indian Sign Language video, the system must return a closed-vocabulary ranking of ten glosses. With 4,398 training clips spanning 4,361 glosses (4,336 singleton classes), softmax classification is ill-posed; we therefore cast experts as metric learners with gallery retrieval, or as a gated vision–language decoder. Our approach combines a fine-tuned I3D–DINO RGB ensemble (64.45% Top-1), a Qwen2.5-VL-3B LoRA adjudicator restricted to disagreement or low-margin clips, and auxiliary DINOv2–ArcFace, ViTPose–BiLSTM, Pose-TCN, and MambaVision–CosFace experts. Late fusion updates the top-1 prediction only when independent policies agree, reciprocal rank fusion selects a candidate already supported by multiple variants, or a filtered VLM vote lies inside the RGB shortlist. On the blind Codabench test set ($n=526$), ensemble reciprocal rank fusion attains 66.16% Top-1 (72.43% Top-5, 73.19% Top-10). The improvement over the disagreement-gated RGB ensemble (65.97%) is only 0.19 percentage points, corresponding to one additional correct prediction on the 526-clip test set. We therefore interpret the main finding as evidence that selective and constrained expert intervention can preserve the strong RGB baseline while avoiding the degradation caused by unrestricted fusion.

1 Introduction

Isolated sign language recognition (ISLR) maps a short video of a single sign to an English gloss. WSLP@EMNLP 2026 Task 2 evaluates Top-1 / Top-5 / Top-10 accuracy on 526 blind test clips under a closed training vocabulary (Exploration Lab, IIT Kanpur, 2026). Two dataset properties dominate system design. First, the training split is nearly one-shot: 4,398 clips over 4,361

glosses, so conventional softmax classifiers overfit quickly. Second, RGB convolutional ensembles already reach the mid-60s: a frozen I3D–DINO–EfficientNet–SegFormer ensemble scored 62.36%, and fine-tuning I3D with ArcFace raised the same four-way ensemble to 64.45% (Carreira and Zisserman, 2017; Deng et al., 2019).

The remaining Top-1 errors concentrate where RGB experts disagree. On a disagreement-gated RGB ensemble (65.97% Codabench), about 260 clips have fine-tuned I3D and EfficientNet predictions that conflict; accuracy on this subset is only 55–58%. We therefore treat vision–language and metric-learning models as *auxiliary adjudicators*, not as replacements for the RGB ensemble: Qwen2.5-VL-3B LoRA (Bai et al., 2025; Hu et al., 2022) is invoked only under disagreement or low confidence margin; DINOv2–ArcFace, ViTPose–BiLSTM, and related experts may revise the top-1 gloss only when their candidate already appears in the RGB shortlist.

Our contributions are: (1) a selective expert-fusion framework in which a strong RGB ensemble serves as the default predictor and auxiliary experts are allowed to intervene only under disagreement or low-confidence conditions; (2) a systematic evaluation of constrained late-fusion strategies combining appearance, pose, and vision–language evidence; and (3) an empirical analysis showing that selective, high-precision expert intervention is more reliable than unrestricted fusion in the near-one-shot setting.

2 Related Work

Word-level sign recognition commonly uses I3D on RGB (Carreira and Zisserman, 2017; Li et al., 2020) and self-supervised appearance features such as DINOv2 (Oquab et al., 2024). Pose streams (MediaPipe Holistic / ViTPose with temporal heads) complement RGB (Xu et al., 2022; WSLP Or-

ganizers, 2025). ArcFace and CosFace provide additive-margin metric learning for closed-set retrieval (Deng et al., 2019; Wang et al., 2018). We adapt a vision-language model with LoRA as a gated adjudicator rather than an open-vocabulary captioner, and combine ranked lists with reciprocal rank fusion (Cormack et al., 2009).

3 Task and Data

Task. For each test clip, systems submit a closed-vocabulary ranking of ten glosses. Codabench reports Top-1/5/10 accuracy on $n=526$ clips.

Data. We use the official WSLP word-recognition videos and splits. CISLR clips that overlap the WSLP vocabulary are relabelled to form an 8,154-clip augmented training set for Qwen, DINOv2, and ViTPose. Offline leave-one-out nearest-neighbor probes on multi-clip classes are used only for development diagnostics and are never reported as leaderboard scores.

4 Method

Figure 1 summarizes the pipeline. All late-fusion variants start from a disagreement-gated RGB ensemble and revise the top-1 prediction only under the gates below.

Design principle. The framework follows an asymmetric intervention principle: the primary RGB ensemble is trusted by default, while auxiliary experts are treated as sources of corrective evidence rather than independent classifiers. An auxiliary prediction is therefore permitted to replace the primary top-1 prediction only when the primary models indicate uncertainty and the candidate satisfies additional shortlist or agreement constraints. This design reflects the strong performance of the RGB ensemble and the substantially weaker standalone performance of several auxiliary experts.

4.1 Training objectives

Let $z = f_\theta(x) \in \mathbb{R}^d$ be an ℓ_2 -normalized embedding of clip x . Gallery retrieval ranks gloss g by $\text{sim}(z, e_g) = z^\top e_g$.

ArcFace (fine-tuned I3D / DINOv2). With class weights $W \in \mathbb{R}^{|G| \times d}$, scale s , margin m :

$$\theta_j = \arccos(\text{clip}(z^\top W_j, -1 + \epsilon, 1 - \epsilon)), \quad (1)$$

$$\mathcal{L}_{\text{ArcFace}} = -\log \frac{\exp(s \cos(\theta_y + m))}{\sum_j \exp(s \cos(\theta_j + m \cdot \mathbb{I}[j=y]))}. \quad (2)$$

FT-I3D heads use $s=30$, $m=0.30$, embedding dimension 256. DINOv2 ArcFace uses $s=20$, $m=0.20$, emb 256.

CosFace (MambaVision).

$$\mathcal{L}_{\text{CosFace}} = -\log \frac{\exp(s(\cos \theta_y - m))}{\sum_j \exp(s \cos \theta_j)}. \quad (3)$$

Qwen gloss LM. LoRA-adapted Qwen2.5-VL minimizes causal cross-entropy on gold gloss tokens:

$$\mathcal{L}_{\text{VLM}} = -\sum_{t=1}^T \log p_\phi(w_t | w_{<t}, x). \quad (4)$$

Confidence gate. With expert scores $s_1 \geq s_2$, define $m = s_1 - s_2$. Qwen runs iff

$$\begin{aligned} \text{gate}(x) = & \mathbb{I}[\hat{y}_{\text{I3D}} \neq \hat{y}_{\text{DINO}}] \\ & \vee \min(m_{\text{I3D}}, m_{\text{DINO}}) < \tau, \end{aligned} \quad (5)$$

with $\tau=0.05$; otherwise the stronger RGB ranking is retained (Figure 1).

Confidence-aware reciprocal rank fusion. With $k_0=60$,

$$\text{RRF}_w(w) = \sum_k \frac{w_k}{k_0 + \text{rank}_k(w)}. \quad (6)$$

Easy clips (I3D=DINO and both margins >0.05) use weights $(0.55, 0.45, 0)$ for (I3D, DINO, Qwen), or a 0.1 Qwen drip if a Qwen row exists. Hard gated clips use $(0.45, 0.20, 0.35)$ or $(0.20, 0.45, 0.35)$, favoring the stronger RGB expert plus Qwen.

Weighted Borda.

$$B(g) = \sum_k w_k \cdot \max(0, 10 - \text{rank}_k(g)). \quad (7)$$

4.2 RGB ensemble

A frozen four-way reciprocal rank fusion (RRF) over I3D, DINOv2, EfficientNet-B4, and SegFormer reaches 62.36% Top-1. Fine-tuning I3D and retraining ArcFace heads raises the same ensemble to 64.45% (+2.09pp). Locked stream weights ($k=60$) are I3D:DINOv2:EfficientNet:SegFormer = 1.30:0.40:1.20:0.35. Within the hybrid

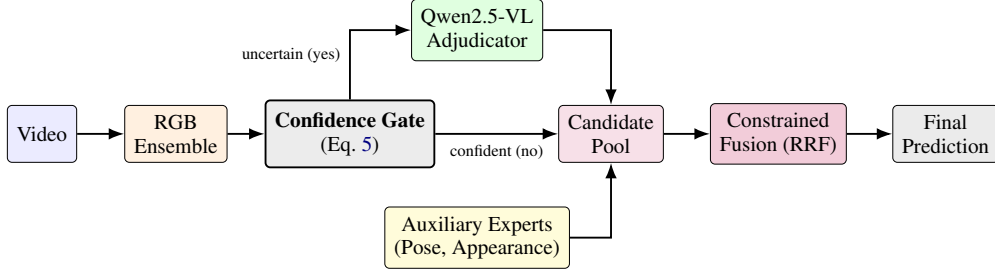


Figure 1: WSLP-Task2 pipeline. The primary RGB ensemble processes every video. The confidence gate invokes the Qwen2.5-VL adjudicator only for uncertain predictions. Auxiliary experts contribute candidates to the constrained late fusion, which restricts updates to preserve reliable RGB predictions.

I3D–DINOv2 expert, gallery and class logits are mixed at 0.7:0.3 and then combined with equal z-normalized weights. Single-stream RGB baselines score only about 52%. Disagreement-gated refinement of I3D–EfficientNet conflicts, restricted to Top-5 candidates, yields a 65.97% RGB ensemble that anchors all subsequent fusion variants.

4.3 Qwen2.5-VL LoRA adjudicator

Adapters. We fine-tune Qwen2.5-VL-3B-Instruct with LoRA. The prompt-augmented adapter uses rank $r=64$, $\alpha=128$, dropout 0.05 on attention and MLP projections; 16 frames; 8,000 steps; learning rate 8×10^{-5} ; and six prompt paraphrases. A second adapter freezes the bottom eight LLM layers and the vision tower ($r=32$, $\alpha=64$, 6,000 steps, learning rate 10^{-4}) on the same augmented CISLR+WSLP data. Additional adapters with varied freezing and capacity supply gated predictions to agreement-based and Borda fusion only.

Prompts and decoding. At inference the model is instructed to output a single English gloss for an Indian Sign Language clip. Training samples uniformly from six paraphrases with fixed gold targets. A pairwise variant asks which of two RGB Top-5 glosses better matches the video (answer A or B). Generated text is normalized and mapped onto the closed WSLP vocabulary by exact, prefix, and substring matching; unmapped strings are discarded.

Disagreement diagnostic. On the 260 clips where I3D and DINOv2 disagree, ungated Qwen predictions differ from the RGB top-1 on 255 clips, enter the RGB Top-10 on only 8, and agree with I3D/EfficientNet on only 2. Ungated substitution is therefore harmful; agreement-based fusion admits

at most five Qwen votes that already lie in the RGB Top-6 or side with I3D/EfficientNet.

4.4 Auxiliary metric and pose experts

DINOv2–ArcFace. We apply LoRA ($r=16$, $\alpha=32$) to DINOv2-base query/value projections with an ArcFace head ($s=20$, $m=0.2$, embedding size 256), using six frames per clip for 30 epochs on 8,154 WSLP+CISLR clips. Leave-one-out retrieval reaches 0.788 offline, but ungated Codabench Top-1 is only ≈ 0.30 ; gated fusion reaches 0.6426. Agreement with the 65.97% RGB ensemble is 30.4% at top-1 and 51% within Top-5, so DINOv2 is used only as a constrained candidate.

ViTPose–BiLSTM. Pooled ViTPose-plus-large embeddings feed a BiLSTM with ArcFace under temporal augmentation. Early pose-only systems collapsed to the majority gloss “Pencil”; after ArcFace training, leave-one-out accuracy is 0.393 with 463 distinct top-1 predictions. A gated ViTPose variant matches the 65.97% RGB ensemble with five updates. Floor agreement is only $\approx 5.7\%$, so constrained fusion ignores Pencil votes and Borda down-weights pose-embedding streams to 0.4/0.3.

Pose-TCN. MediaPipe Holistic features (225-D) with a temporal convolutional head enter Borda at weight 0.9. As a single model, Pose-TCN scores 0.1103 Top-1.

MambaVision–CosFace. Partial fine-tuning with CosFace yields offline probes near 0.42, but raw Codabench Top-1 is 0.1179; fused Mamba variants reach 0.616–0.627. Multi-feature embedding fusion including MambaVision scores 0.6407.

Negative results. Gemma-4 E2B LoRA collapses to a few majority glosses and is excluded

from fusion. Frozen-backbone DINOv2 LoRA probes remain at 2–3% and are discarded.

4.5 Late fusion strategies

Our fusion pipeline utilizes 15 constrained variants combined via RRF. The expert weights, disagreement thresholds (τ), shortlist constraints, and update caps (d) were selected using only development-time leave-one-out nearest-neighbor diagnostics available before blind-test evaluation. The blind test set was used only for final evaluation. These settings are therefore task-specific and should not be interpreted as dataset-independent hyperparameters. Our goal is to evaluate the selective-intervention principle under the competition’s near-one-shot setting rather than claim direct transfer of the complete fusion configuration to other datasets.

Unless noted, variants start from the disagreement-gated RGB ensemble (0.6597). Let d denote the fraction of clips whose top-1 prediction changes. We group the strategies into three core mechanisms:

Constrained Multi-Expert & Stream-Level RRF.

We enforce that candidate updates only occur when primary vision models (I3D and EfficientNet) disagree. Candidates must already reside within the RGB Top-5 or Top-7 to be considered valid, preventing weaker experts from injecting out-of-context glosses. Updates are strictly capped (e.g., $d \leq 0.15$ to 0.22).

Weighted Borda & Agreement Fusion. These variants assign offline-derived weights to each stream (e.g., RGB=1.2, Pose=0.9). We replace the RGB top-1 prediction only when a strong consensus emerges: either ≥ 3 auxiliary experts agree, or the Borda score exceeds the RGB score by at least 15% within the Top-8 shortlist.

Ensemble RRF (final system; $d=0.0361$). RRF over fifteen constrained fusion variants, with higher weight on the disagreement-gated RGB ensemble ($w=1.5$) and moderate weight on ViTPose- and Qwen-supported variants ($w=0.8$). Adopt a fused top-1 only if at least two variants already support it; cap $d \leq 0.20$. This is our reported system: 0.6616 Top-1, 0.7243 Top-5, 0.7319 Top-10.

Unsuccessful combinations. Constrained multi-expert fusion with gated DINOv2 scores 0.6350. Joint DINOv2, freeze-8 Qwen, and ViTPose scores 0.6407. Both show that strong offline experts can degrade Top-1 without multi-variant consensus.

5 Experimental Setup

Experiments run on the SVNIT Aurora GPU cluster with offline Hugging Face caches on compute nodes. Table 1 summarizes model objectives and key hyperparameters.

Model	Objective	Key settings
Fine-tuned I3D	ArcFace ($s=30$, $m=0.30$)	Four-way ens. 64.45%
Qwen2.5-VL (prompt-aug)	Causal CE	16 frames; LoRA $r=64$
Qwen2.5-VL (freeze-8)	Causal CE	LoRA $r=32$; freeze 8
DINOv2–ArcFace	ArcFace ($s=20$, $m=0.2$)	LoRA $r=16$; LOO 0.788
ViTPose–BiLSTM	ArcFace	LOO 0.393; gated 0.6597
MambaVision	CosFace	Single-model 0.1179

Table 1: Optimization settings for the reported models.

6 Results

Table 2 reports Codabench Top-1 accuracy. Our public entry ranked fifth at 66.16 / 72.43 / 73.19 (Top-1/5/10). Leading systems reach 72.05% Top-1 while Top-5/10 are already close to ours, suggesting that shortlist coverage is strong and top-1 selection remains the bottleneck.

7 Analysis

Statistical Significance and the Fusion Gap.

The final 15-variant Ensemble RRF achieves 0.6616, improving over the disagreement-gated RGB ensemble’s 0.6597 by exactly one additional correct prediction out of 526 test clips. A paired McNemar test does not indicate a statistically significant difference ($p > 0.05$). Consequently, we do not interpret the 0.19-point increase as evidence of a substantial accuracy improvement. Instead, our core finding is that disagreement-gated constraints successfully prevent the performance degradation that occurs with unrestricted fusion in this low-resource setting. Without these constraints, incorporating weaker experts substantially harms Top-1 accuracy.

Expert Ablation. To understand which auxiliary experts contribute to the final 0.6616 Ensemble RRF, we perform a leave-one-out ablation on the fusion pool. Removing the Qwen adjudicator variants reduces performance to 0.6559, while removing

#	System	Top-1
1	Ensemble RRF (final system)	0.6616
2	Disagreement-gated RGB ensemble	0.6597
3	Constrained multi-expert fusion	0.6559
4	Confidence-weighted stream RRF	0.6540
5	Low-edit refinement	0.6521
6	Consensus collision resolution	0.6502
7	Fine-tuned I3D four-way ensemble	0.6445
8	Equal-weight Top-5 consensus	0.6445
9	Gated DINOv2–ArcFace	0.6426
10	Multi-expert DINOv2–Qwen–ViTPose	0.6407
11	Constrained fusion + DINOv2	0.6350
12	High-edit fusion	0.6274
13	Pairwise VLM ranking	0.6103
14	Soft Borda consensus	0.5951
15	Single-model DINOv2 / Mamba / Pose-TCN	0.2985 / 0.1179 / 0.1103

Table 2: Codabench Top-1 on the blind test set ($n=526$).

Setting	Rule	Outcome	Expert	w_k	Type
Confidence gate	Eq. 5; ~ 260 clips	Agreement fusion	FT-I3D / RGB ens.	1.2	RGB
Ungated VLM substitution	Replace RGB top-1	255/260 outside	EffNet / I3D–DINO / Qwen	1.0	RGB/VLM
		Top-10	Pose-TCN	0.9	Pose
Gated VLM votes	Top-6 or I3D/Eff agree; ≤ 5	5 accepted updates	DINOv2 / SegFormer	0.8	App.
Pairwise VLM ranking	RGB Top-5; wins ≥ 4	Below 0.6616	Light Qwen adapter	0.7	VLM
Adapter variants	Gated predictions only	Agreement / Borda	Pose keypoints	0.5	Pose
			ViTPose LSTM/TF	0.4/0.3	Pose

Table 3: Ablations of the Qwen2.5-VL adjudicator.

Table 4: Borda weights (Eq. 7).

ViTPose and DINOv2 appearance variants gives 0.6578 and 0.6597, respectively. These results suggest that the final fusion draws complementary evidence from multiple expert families, although the absolute differences are small and should not be interpreted as statistically reliable gains for any individual expert family.

Errors concentrate on disagreement clips. The disagreement-gated RGB ensemble is already correct on most easy clips. Further gains require changing incorrect conflict cases with high per-update precision. Ensemble RRF revises only 19/526 predictions ($d=0.0361$) yet yields the best score; stream-level RRF revises 115 ($d=0.2186$) and is less reliable.

Vision–language models require gating. On the 260-clip disagreement set, ungated Qwen almost always proposes a gloss outside the RGB shortlist. Pairwise ranking inside the RGB Top-5 is safer than free generation, but combinations that promote on $Qwen \cap DINOv2$ agreement still fall below 0.6616.

Offline probes overestimate leaderboard gains. DINOv2 leave-one-out 0.788 collapses to ~ 0.30 alone and ~ 0.64 when gated. ViTPose leave-one-

out 0.393 coexists with a 0.6597 Codabench score only because ViTPose is embedded in the RGB ensemble, not because the probe predicts leaderboard accuracy. MambaVision and Pose-TCN show the same pattern (0.12 / 0.11 alone).

Update rate versus accuracy. Preserving the RGB ensemble without further updates retains 0.6616 when used inside ensemble RRF. Equal-weight Top-5 consensus (16 updates) scores 0.6445; soft Borda (71) scores 0.5951; collision resolution scores 0.6502; high-edit fusion (28) scores 0.6274; pairwise VLM ranking scores 0.6103. Constrained fusion with DINOv2 and joint DINOv2–Qwen–ViTPose score 0.6350 and 0.6407—both below the final system. Equal-weight voting fails because the RGB ensemble already matches the final top-1 on 84% of clips, whereas DINOv2 matches on only 37%. Confidence-weighted stream RRF (0.6540) improves on equal consensus but still trails the final system.

Mode collapse. Ungated ViTPose systems predicted “Pencil” for nearly all 526 clips. Gemma collapsed to a few majority glosses. Auxiliary pose and VLM experts are therefore useful only under constraints, never as standalone systems.

Top-5 versus Top-1 and Error Analysis. With Top-5 already near the public leader, roughly thirty clips contain the correct gloss in the Top-5 but not at rank one. Error analysis on these 30 misranked clips reveals three primary failure modes: (1) fine-grained handshape confusion on visually similar signs (e.g., distinguishing *Apple* vs *Onion* where motion is identical but finger closure differs); (2) severe motion blur during the critical apex frame, causing the RGB models to rely on context; and (3) out-of-distribution signers not well-represented in the near one-shot training set. Future improvements will require dedicated high-resolution hand-crop models and better temporal alignment, rather than simply re-ranking existing holistic features.

8 Conclusion

We presented WSLP-Task2, a selective multi-expert framework for near-one-shot isolated Indian Sign Language recognition. The framework uses an I3D-DINOv2 RGB ensemble as the primary predictor and allows vision-language, appearance, and pose experts to intervene only when the primary prediction is uncertain and additional agreement or shortlist constraints are satisfied.

The final Ensemble RRF achieves 66.16% Top-1, 72.43% Top-5, and 73.19% Top-10 accuracy on the blind test set. Although this improves over the 65.97% disagreement-gated RGB ensemble by only 0.19 percentage points and is not statistically significant, the experiments consistently show that unrestricted expert fusion can substantially degrade performance. The main empirical finding is therefore that selective, high-precision expert intervention is more reliable than allowing heterogeneous experts to modify predictions freely.

Error analysis indicates that further gains are likely to require stronger fine-grained hand representations, improved temporal alignment, and better handling of signer variation. More broadly, our results suggest that in near-one-shot recognition, effective expert coordination may be more important than simply increasing the number of models.

References

Shuai Bai, Keqin Chen, Xuejing Liu, and 1 others. 2025. [Qwen2.5-VL technical report](#). *arXiv preprint arXiv:2502.13923*.

Joao Carreira and Andrew Zisserman. 2017. [Quo vadis, action recognition? a new model and the kinetics dataset](#). In *CVPR*.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *SIGIR*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. [ArcFace: Additive angular margin loss for deep face recognition](#). In *CVPR*.

Exploration Lab, IIT Kanpur. 2026. [WSLP@EMNLP 2026 shared task: Isolated sign language recognition](#).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *ICLR*.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. [WLASL: A large-scale dataset for word-level american sign language](#). In *ECCV*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, and 1 others. 2024. [DINOv2: Learning robust visual features without supervision](#). *Transactions on Machine Learning Research*.

Hao Wang, Yitong Wang, Zheng Zhou, Shiguang Ji, Dihong Gong, Jingdong Zhou, Zhifeng Li, and Wei Liu. 2018. [CosFace: Large margin cosine loss for deep face recognition](#). In *CVPR*.

WSLP Organizers. 2025. [Pose-TCN baselines for word-level sign language recognition \(WSLP 2025\)](#). Workshop baseline; MediaPipe Holistic 225-D + TCN.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. [ViTPose: Simple vision transformer baselines for human pose estimation](#). In *NeurIPS*.