The CUHKSZ System for the IWSLT 2026 Low-Resource Speech-to-Text Task

Ruiyan Sun[1*] **Qingming Li**[1*] **Satoshi Nakamura**[2†]
[1] School of Data Science, The Chinese University of Hong Kong, Shenzhen
[2] School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen
ruiyansun@link.cuhk.edu.cn qingmingli@link.cuhk.edu.cn nakamura@ai.cuhk.edu.cn

Abstract

This paper describes the CUHKSZ system for the IWSLT 2026 Low-Resource Speech-to-Text task. We propose Gradient-Driven Parameter Sharing (GDPS), a framework that analyzes inter-language gradient behaviors to automatically determine optimal language groupings and shared-private parameter ratios. Built upon SeamlessM4T-Medium, GDPS reduces negative transfer by specializing Layer 11 FFN2 while maintaining shared encoder representations across languages. Additionally, we incorporate curriculum distillation with progressive pseudo-label mixing and test-time reranking combining prior-BLEU weighting and self-consistency scoring. Evaluation on eight low-resource languages (bem, ckb, gle, hau, ibo, yor, aeb, est) demonstrates strongest gains on bem (+2.07 BLEU), hau (+1.50), and ibo (+0.38) compared to unified fine-tuning, while ckb and yor benefit more from prior-based reranking at inference.

1 Introduction

The IWSLT 2026 Low-Resource Speech-to-Text (S2TT) track (Adelani et al., 2026) focuses on multilingual speech translation for languages with limited training data. This task presents significant challenges due to the scarcity of parallel speech-text corpora and the linguistic diversity across target languages. We participate in six of the official language directions (bem, ckb, gle, hau, ibo, yor) and additionally include two prior-edition directions (aeb, est), yielding an 8-language setup. All systems are submitted under the unconstrained condition (external pre-trained models and prior-edition data are used).

Massively multilingual models such as SeamlessM4T (SEAMLESS Communication Team, 2025; Barrault et al., 2023b), Seamless v2 (Barrault et al., 2023a), and MMS (Pratap et al., 2024) show strong cross-lingual transfer. However, when fine-tuning for multiple low-resource languages simultaneously, language-specific gradient signals often conflict during backpropagation, causing performance degradation for underrepresented languages (Yu et al., 2020)—a phenomenon known as *negative transfer*.

Gradient conflict is a primary bottleneck in multilingual learning. Prior work mitigates it along three lines. (1) *Optimization-based*: PCGrad (Yu et al., 2020) projects conflicting gradients onto normal planes; GDOD (Dong et al., 2022) and GradOPS (Zhu et al., 2025) decompose gradients into shared/task-specific subspaces via orthogonal projection. These stabilize optimization but keep a single shared parameter set. (2) *Architectural*: shared-private representations (Bousmalis et al., 2016) and orthogonal LoRA disentangling (Yang et al., 2026) decouple interfering parameters, but the partition is typically hand-designed. (3) *MoE*: routes tokens to specialized modules at the cost of substantially more parameters. GDPS differs from all three: it *automatically* determines both the language grouping and the shared-private ratio from measured gradient statistics, instantiating the decision as a localized FFN2 decomposition. This builds on our IWSLT 2025 system (Sun and Nakamura, 2026) (fixed sharing) and differs from other IWSLT low-resource systems that rely on data augmentation or larger backbones (Meng and Anastasopoulos, 2025; Robinson et al., 2025).

Concretely, GDPS targets the FFN2 sublayer in Conformer encoders because it has the highest parameter density among sublayers and is more malleable for language-specific features than attention modules (Gerber, 2025). We specialize only Layer 11 FFN2—the bottleneck identified by our gradient analysis—minimizing architectural changes while maximizing language-specific adaptation.

[*] Both authors contributed equally.
[†] Corresponding author.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 296–304

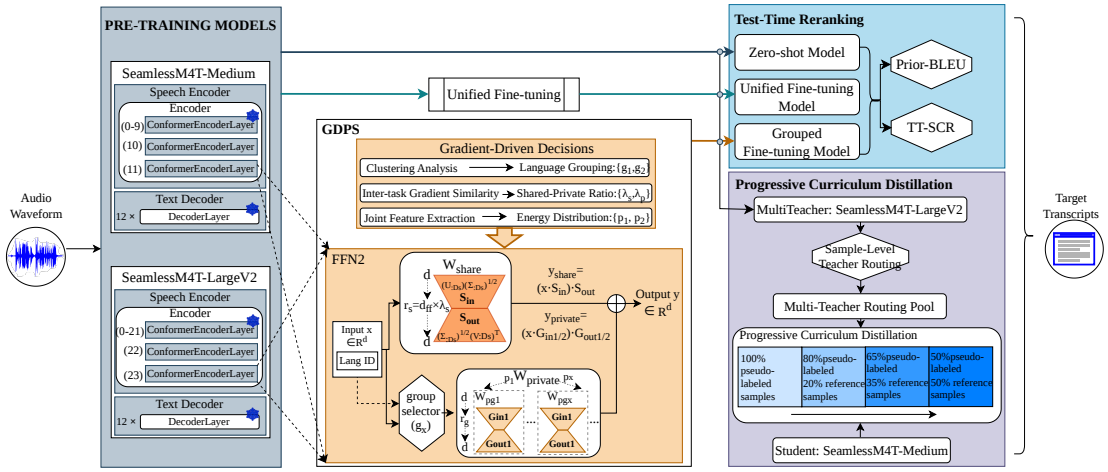

Figure 1: GDPS system overview with four main stages: (1) backbone model pre-training; (2) gradient conflict analysis via three methods to identify language groups and layer bottlenecks; (3) targeted FFN2 specialization with shared-private decomposition; (4) curriculum distillation training and test-time reranking.

Our system builds upon SeamlessM4T-Medium and incorporates curriculum distillation using SeamlessM4T-v2-Large as teacher to bridge the gap between pretrained knowledge and low-resource fine-tuning. At inference time, we employ dual-strategy test-time reranking combining prior-BLEU weighting and test-time self-consistency reranking (TT-SCR) to select the best translation from multiple heterogeneous checkpoints.

This paper describes our system architecture, training pipeline, and experimental setup for the IWSLT 2026 evaluation. Our main contributions are:

- An automated gradient-driven framework that translates gradient statistics into optimal GDPS configurations for 8 languages
- A curriculum distillation protocol that stabilizes training through progressive data incorporation
- A dual-strategy test-time reranking framework for robust inference across heterogeneous checkpoints
- Comprehensive evaluation on 8 languages demonstrating where GDPS yields gains and where UFT remains stronger

2 System Description

2.1 Pre-trained Model

We adopt SeamlessM4T-Medium (Barrault et al., 2023b) as our backbone model, which is a massively multilingual speech-to-text model supporting up to 100 languages. The model consists of a speech encoder (12 Conformer layers) and a text decoder (12 layers), with approximately 1.2B parameters.

2.2 Data Processing

We conduct experiments on the IWSLT 2026 Low-resource Speech-to-Text track, covering two language sets:

- **8-Language Setup:** Six IWSLT 2026 official directions—Bemba (bem), Central Kurdish (ckb), Irish (gle), Hausa (hau), Igbo (ibo), Yoruba (yor)—plus two prior-edition directions: Tunisian Arabic (aeb, IWSLT 2022) and Estonian (est, LoResMT). All submitted under the unconstrained condition.
- **6-Language Subset (IWSLT 2026 Official):** bem, ckb, gle, hau, ibo, yor. Models trained only on the 2026-provided data for these six directions.

Dataset statistics are summarized in Table 1.

All speech inputs are preprocessed at 16kHz following IWSLT 2026 protocols (lowercase, no punctuation).

2.3 Training Strategy

Our training pipeline consists of three main stages:

2.3.1 Unified Fine-tuning Baseline

We first establish a unified fine-tuning baseline where all languages are jointly fine-tuned with

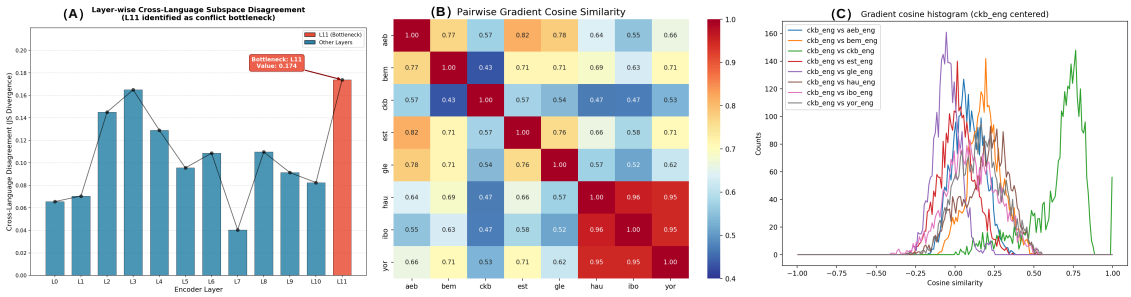

Figure 2: Gradient conflict analysis identifying bottlenecks and language groupings: (A) Layer-wise subspace disagreement reveals Layer 11 as the primary conflict point; (B) Pairwise gradient similarity shows ckb as a clear outlier with distinct cross-lingual patterns; (C) Gradient cosine histogram centered on ckb_eng, showing near-zero cross-language similarity versus high self-similarity, confirming ckb's outlier status.

Table 1: Dataset summary. The 6-language subset corresponds to the IWSLT 2026 official directions; the 8-language setup additionally includes prior-edition data (aeb, est).

Language	Task	Amount	Sources
Tunisian (aeb)	2-way ST	20k lines	IWSLT2022 (Anastasopoulos et al., 2022)
Bemba (bem)	2-way ST	20k lines	BIG-C (Sikasote et al., 2023)
Central Kurdish (ckb)	2-way ST	9k lines	IWSLT 2026
Estonian (est)	2-way ST	20k lines	LoResMT (Sildam et al., 2024)
Irish (gle)	2-way ST	10k lines	IWSLT2023 (Agarwal et al., 2023)
Hausa (hau)	2-way ST	15k lines	IWSLT 2026
Yoruba (yor)	2-way ST	15k lines	IWSLT 2026
Igbo (ibo)	2-way ST	14k lines	IWSLT 2026

uniform parameter sharing. This captures cross-lingual transfer through shared weights and serves as the starting point for GDPS specialization.

Key training hyperparameters are summarized in Table 2.

Table 2: Training configuration summary.

Parameter	Value
Base Learning Rate (α)	4×10^{-5}
Group-Adjusted Learning Rate (α_g)	1×10^{-4}
Batch Size (B)	4
Dropout (p_d)	0.05
Weight Decay (λ)	0.05
Warmup Steps (t_w)	2000
Seed (s)	2343
Optimizer	AdamW
Precision	FP16
Hardware	NVIDIA A100

Training uses mixed-precision (FP16) with AdamW optimizer on NVIDIA A100 GPUs.

2.3.2 Gradient-Driven Parameter Sharing (GDPS)

GDPS (Sun and Nakamura, 2026) automatically discovers the optimal parameter sharing configuration by analyzing inter-language gradient behaviors. The framework consists of three components:

Method A - Language Grouping via Clustering: We compute pairwise gradient cosine similarity between languages at the layer level. Let $\bar{g}_i \in \mathbb{R}^d$ denote the averaged gradient vector of language i across all samples at a given layer. The pairwise cosine similarity is:

$$s_{i,j} = \frac{\bar{g}_i \cdot \bar{g}_j}{\|\bar{g}_i\| \|\bar{g}_j\|} \tag{1}$$

We convert similarity to distance $d_{i,j} = 1 - s_{i,j}$ and apply hierarchical clustering with Ward linkage to partition languages into groups sharing parameters. As illustrated in Figure 2(B), the 8×8 pairwise similarity heatmap reveals distinct language clusters.

Method B - Self versus Cross Gradient Similarity: To identify which languages are pulling parameters in conflicting directions, we compute intra-task and inter-task gradient similarity. Let $\mathbf{g}_{t,i}$ denote the gradient vector for language i at a sample t. We compute:

$$S_{\text{self}} = \mathbb{E}_{i \neq j} \mathbb{E}_t \cos(\mathbf{g}_{t,i}, \mathbf{g}_{t,j}) \tag{2}$$

$$S_{\text{cross}} = \mathbb{E}_{i,j} \mathbb{E}_{t \neq t'} \cos(\mathbf{g}_{t,i}, \mathbf{g}_{t',j}) \tag{3}$$

The conflict strength is $\delta = S_{\text{self}} - S_{\text{cross}}$: high values indicate that language pairs agree strongly within batches but diverge across batches, suggesting batch-level gradient conflicts. This score maps to a shared ratio:

$$\text{SharedRatio} = \begin{cases} 0.75 & \delta < 0.05 \\ 0.50 & 0.05 \leq \delta < 0.15 \\ 0.25 & \delta \geq 0.15 \end{cases} \tag{4}$$

Figure 2(C) presents the sample-level analysis, where ckb exhibits notably higher self-similarity (0.509) compared to cross-language pairs ($\sim$0.21), confirming its outlier status.

Method C - Joint SVD and Regularized CCA: Let $\mathbf{G}_i \in \mathbb{R}^{m \times d}$ be the gradient matrix of language i (stacking sample-level gradients across m samples with d dimensions). We concatenate all language gradient matrices as $\mathbf{G}_{\text{concat}} = [\mathbf{G}_1; \ldots; \mathbf{G}_n] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ via SVD. Ridge-regularized CCA (Hotelling, 1936; Hardoon et al., 2004) maximizes cross-covariance between language pairs. The energy concentration for language i is $p_i = E_i / \sum_{l=1}^{n} E_l$, where $E_i = \sum_{j=1}^{k} \|\mathbf{G}_i \mathbf{v}_j\|^2$ measures the cumulative projection energy onto the top-k singular vectors.

Layer Selection: As shown in Figure 2(A), cross-language subspace disagreement computed via CCA across all 12 encoder layers reveals that Layer 11 FFN2 exhibits the highest disagreement (0.174), confirming it as the primary gradient conflict bottleneck. This layer is selected for FFN2 specialization.

Language Grouping Results: Hierarchical clustering with Ward linkage on the 8×8 distance matrix is applied independently per backbone, as larger models exhibit finer gradient separation. For **SeamlessM4T-Medium** (K=3), the dendrogram yields three clusters:

- Group 0: ckb (Central Kurdish) — isolated cluster with weak connectivity (silhouette 0.33), indicating its gradient profile differs substantially from all other languages
- Group 1: hau, ibo, yor (Niger-Congo family) — tight cluster (silhouette ∼0.55) reflecting high mutual gradient similarity
- Group 2: aeb, bem, est, gle (European/Afro-Asiatic) — cohesive cluster (silhouette ∼0.48) with moderate shared gradients

For **SeamlessM4T-v2-Large** (K=4), the higher model capacity resolves a finer cluster boundary, yielding four groups: G0 {bem, hau, ibo, yor} (Niger-Congo and Afro-Asiatic low-resource directions with high mutual gradient similarity), G1 {aeb, est} (higher-resource directions with more independent gradient profiles), G2 {gle} (Irish, isolated due to near-zero translation performance and highly concentrated gradient energy), and G3 {ckb} (high-conflict outlier, same as the Medium setting).

Shared-Private Ratio Results: The conflict score $\delta \approx 0.08$ falls in the medium conflict range ($0.05 \leq \delta < 0.15$), yielding a 50% shared ratio. However, ckb exhibits $\delta > 0.15$, indicating it would benefit from a lower shared ratio.

Energy Distribution: Joint SVD shows high gradient concentration (Gini coefficients 0.74–0.85), with gle having the highest concentration (0.85) and ckb the lowest (0.74), reflecting ckb's isolated gradient profile.

Figure 2 presents the complete gradient conflict analysis visualization.

Architecture Instantiation: We specialize Layer 11 FFN2 by decomposing $\mathbf{W}_{\text{unified}} = \mathbf{W}_2\mathbf{W}_1$ into shared and group-specific components. Let d_{model}=1024, d_{ff}=4096, and shared ratio ρ=0.5. We compute the equivalent Gram matrix $\mathbf{W}_{\text{equiv}} = (\mathbf{W}_2\mathbf{W}_1)^\top$ and apply SVD: $\mathbf{W}_{\text{equiv}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Top-$k$ (k=$d_{\text{model}} \times \rho$=512) singular vectors capture universal structure; the shared component is $\mathbf{W}_{s,1} = \mathbf{U}_{:,1:k}\sqrt{\mathbf{\Sigma}_{1:k,1:k}}$, $\mathbf{W}_{s,2} = \sqrt{\mathbf{\Sigma}_{1:k,1:k}}\mathbf{V}^\top_{:,1:k}$, expanded to r_s=$d_{\text{ff}} \times \rho$=2048. The remaining capacity (d_{pv}=2048) is split across N language groups (N=3 for Medium, N=4 for LargeV2) proportional to their energy concentration β_g. The final layer combines shared and private paths:

$$\text{FFN2}(x) = \text{SiLU}(x\widehat{\mathbf{W}}_{s,1})\widehat{\mathbf{W}}_{s,2} \oplus \bigoplus_{g=1}^{N} \text{SiLU}(x\mathbf{W}_{g,1})\mathbf{W}_{g,2} \quad (5)$$

where $\oplus$ denotes feature concatenation. This maintains universal representations while allowing group-specific adaptation.

2.3.3 Curriculum Distillation

For enhanced training stability, we apply curriculum distillation (Hinton et al., 2015) using SeamlessM4T-v2-Large as the teacher model. The protocol consists of four stages: **Stage 0 (Distill)** uses teacher-generated pseudo-labels; **Stage 1 (mix20)** uses 20% real data + 80% pseudo labels; **Stage 2 (mix35)** uses 35% real data + 65% pseudo labels; **Stage 3 (mix50)** uses 50% real data + 50% pseudo labels.

2.4 Inference Strategy

At inference time, we have multiple heterogeneous checkpoints: SeamlessM4T-Medium zero-shot, Unified FT checkpoints, and GDPS checkpoints. No single checkpoint uniformly dominates all languages. We propose a test-time reranking framework with two complementary strategies.

Let $\mathcal{M}$ denote the set of available model checkpoints, each producing a candidate translation y_m^i for sample i.

Table 3: BLEU/chrF++ on 8-language set. ZS=Zero-shot, UFT=Unified Fine-tuning, GDPS=Our method. Languages are grouped into lower-conflict and the high-conflict outlier (ckb) identified by the gradient analysis. Avg = macro-average across all 8 languages.

	Lower-conflict languages							High-confl.	
Method	**aeb**	**bem**	**est**	**gle**	**hau**	**ibo**	**yor**	**ckb**	**Avg**
ZS-Med	3.58/17.48	0.87/15.58	30.44/53.74	0.07/6.01	0.50/11.49	0.74/10.38	7.50/27.44	7.19/22.97	6.36/20.64
ZS-LargeV2	4.02/18.21	0.81/13.72	29.29/50.63	0.06/8.48	0.59/11.68	0.97/11.11	12.06/33.86	**19.23**/40.96	8.38/23.58
UFT-Med	7.70/24.19	18.49/41.58	40.06/62.04	0.10/10.38	1.80/20.64	3.00/20.44	8.43/29.39	10.67/29.17	11.28/29.73
GDPS-Med	8.42/25.79	20.56/43.70	39.09/61.35	0.16/10.14	3.30/23.38	3.38/20.76	8.03/29.20	8.38/26.46	11.41/30.10
UFT-LV2 (ref.)	**8.66/26.79**	18.32/39.78	**47.96/68.32**	0.08/11.57	1.58/20.11	1.96/18.87	**13.20/35.07**	19.03/**40.65**	**13.85/32.64**
GDPS-LV2 (ref.)	8.54/26.31	18.27/40.67	47.04/67.29	0.07/10.33	1.82/20.31	1.98/18.81	11.91/34.00	17.35/38.59	13.37/32.04

Table 4: BLEU/chrF++: 8L vs 7L (ckb removed). ckb=N/A in the 7L setting. 8L rows show absolute BLEU/chrF++; **each 7L cell reports the deltas ΔBLEU/ΔchrF++ relative to the corresponding 8L row**, isolating the effect of excluding the high-conflict outlier ckb. Each part is independently colored (red = gain, blue = drop). The benefit is capacity-dependent: consistent for GDPS-LV2 (6/7 languages up on BLEU) but small and mixed for the other settings.

	Lower-conflict languages							High-confl.	
Model	**aeb**	**bem**	**est**	**gle**	**hau**	**ibo**	**yor**	**ckb**	**Avg**
8L-UFT-Med	7.70/24.19	18.49/41.58	40.06/62.04	0.10/10.38	1.80/20.64	3.00/20.44	8.43/29.39	10.67/29.17	11.28/29.73
8L-GDPS-Med	8.42/25.79	20.56/43.70	39.09/61.35	0.16/10.14	3.30/23.38	3.38/20.76	8.03/29.20	8.38/26.46	11.41/30.10
8L-UFT-LV2	8.66/26.79	18.32/39.78	47.96/68.32	0.08/11.57	1.58/20.11	1.96/18.87	13.20/35.07	19.03/40.65	13.85/32.64
8L-GDPS-LV2	8.54/26.31	18.27/40.67	47.04/67.29	0.07/10.33	1.82/20.31	1.98/18.81	11.91/34.00	17.35/38.59	13.37/32.04
7L-UFT-Med	−0.08/+0.33	+0.42/+0.26	+0.34/+0.18	−0.01/−0.13	+0.22/+0.52	+0.20/+0.18	−0.62/−0.30	N/A	+0.16/+0.23
7L-GDPS-Med	+0.19/+0.28	−0.03/−1.06	+0.30/+0.08	−0.09/−0.40	+0.26/+0.31	−0.04/−0.24	−0.09/−0.72	N/A	+0.51/+0.27
7L-UFT-LV2	+0.01/−0.37	−0.98/−0.09	+0.70/+0.28	+0.03/−0.52	−0.07/−0.56	−0.06/+0.14	−0.25/−0.08	N/A	−0.83/−1.31
7L-GDPS-LV2	+0.85/+2.04	+1.30/+1.19	+0.96/+0.90	+0.05/−0.87	+0.48/+1.25	+0.33/+0.63	−0.34/−0.58	N/A	−0.05/−0.29

Prior-BLEU + Consensus Selection: We assign each model a prior reliability score R_m estimated from historical validation BLEU. For each sample, we select the model m^* that maximizes:

$$m^* = \text{argmax}_{m \in \mathcal{M}} \big[\alpha \cdot \tilde{R}_m + \beta \cdot \frac{1}{|\mathcal{M}|-1} \sum_{n \neq m} \text{chrF}(y_m^i, y_n^i) - \gamma \cdot P(y_m^i)\big] \quad (6)$$

where $\tilde{R}_m$ is the min-max normalized prior score, the second (consensus) term measures minimum-Bayes-risk-style agreement with other model hypotheses, and $P(\cdot)$ is a degeneration penalty. In our official submission we use chrF++ as the consensus metric with weights $(\alpha, \beta, \gamma) = (0.6, 0.3, 0.1)$, tuned on a validation set; when the top-two candidates differ by less than a margin of 0.05 the selection falls back to the candidate with the highest prior, which stabilizes low-confidence cases. We refer to this full configuration as *Prior-BLEU+Consensus*; setting $\beta = 0$ recovers the prior-only ablation we report as *Prior-BLEU*.

The prior scores $\tilde{R}_m$ encode cross-lingual transfer efficiency: unified fine-tuned models score higher on languages with abundant in-domain data, while GDPS models score higher on languages exhibiting strong gradient conflict during training.

Test-Time Self-Consistency Reranking (TT-SCR): TT-SCR is a prior-free alternative that relies entirely on test-time signals:

$$\text{TT-SCR}(m, i) = \frac{1}{|\mathcal{M}|-1} \sum_{n \neq m} \text{chrF}(y_m^i, y_n^i) - \lambda \cdot P(y_m^i) \quad (7)$$

The selection is $m^* = \text{argmax}_m$ TT-SCR(m, i).

TT-SCR exploits the insight that models producing translations agreeing with many other models tend to be better calibrated. To identify challenging samples, we compute consensus disagreement as $\Delta_i = \max_{m,n} \text{chrF}(y_m^i, y_n^i) - \min_{m,n} \text{chrF}(y_m^i, y_n^i)$. High disagreement indicates model uncertainty, which flags hard samples.

3 Experiments

3.1 Evaluation Setup

Evaluation follows the IWSLT 2026 official protocol: standard lowercase BLEU (sacreBLEU) without punctuation. We report BLEU and chrF++ (Popović, 2015) on the validation set as primary metrics.

3.2 Main Results

Table 3 presents BLEU/chrF++ results for all 8 language pairs.

Table 5: BLEU/chrF++ on curriculum distillation and test-time reranking. Prior-BLEU+Consensus is our official submission configuration; Prior-BLEU is its prior-only (β=0) ablation. Avg = macro-average across languages.

	Lower-conflict languages							High-confl.	
Stage/Strategy	**aeb**	**bem**	**est**	**gle**	**hau**	**ibo**	**yor**	**ckb**	**Avg**
GDPS (baseline)	8.42/25.79	20.56/43.70	39.09/61.35	**0.16/10.14**	3.30/23.38	3.38/20.76	8.03/29.20	8.38/26.46	11.41/30.10
Distill-only	6.92/22.00	17.27/39.74	36.76/58.88	0.10/8.13	1.28/16.03	2.32/16.48	6.19/25.48	3.50/15.78	9.29/25.31
mix20	7.01/22.59	17.92/40.78	37.74/59.82	0.08/7.56	1.09/15.96	1.97/15.61	5.57/24.58	4.05/16.93	9.43/25.48
mix35	7.41/23.27	18.93/41.98	38.82/60.73	0.09/7.90	1.30/16.80	2.42/16.23	6.50/25.94	3.84/16.92	9.91/26.22
mix50	7.60/23.31	19.23/41.73	38.42/60.40	0.07/7.97	1.52/17.47	2.44/17.07	6.30/25.66	4.50/17.87	10.01/26.43
Prior-BLEU	9.43/28.35	**20.60/43.84**	**48.07/68.33**	0.12/9.59	**3.56/23.69**	3.38/20.76	13.20/35.07	19.23/40.96	14.70/33.82
Prior-BLEU+Consensus	**9.62/28.65**	20.56/43.71	48.00/68.21	0.12/9.46	**3.56/23.69**	**3.43/20.93**	**13.23/35.11**	**20.26/42.09**	**14.85/33.98**
TT-SCR	8.10/25.89	18.70/43.58	40.47/62.54	0.09/10.92	2.41/22.26	3.26/21.22	8.64/30.58	10.24/29.72	11.49/30.84

Table 6: BLEU/COMET on 8-language set. Avg = macro-average across languages.

	Lower-conflict languages							High-confl.	
Method	**aeb**	**bem**	**est**	**gle**	**hau**	**ibo**	**yor**	**ckb**	**Avg**
ZS-Med	3.58/**0.513**	0.87/0.414	30.44/0.726	0.07/0.379	0.50/0.384	0.74/0.395	7.50/0.565	7.19/0.510	6.36/0.486
ZS-LargeV2	4.02/0.510	0.81/0.402	29.29/0.737	0.06/0.395	0.59/0.389	0.97/0.401	12.06/0.628	**19.23/0.661**	8.38/0.515
UFT-Med	7.70/0.554	18.49/0.688	40.06/0.758	0.10/0.409	1.80/0.492	3.00/0.496	8.43/0.596	10.67/0.553	11.28/0.568
GDPS-Med	8.42/0.567	**20.56/0.707**	39.09/0.753	**0.16/0.419**	**3.30/0.524**	**3.38/0.498**	8.03/0.590	8.38/0.527	**11.41/0.573**
UFT-LV2 (ref.)	**8.66/0.571**	18.32/0.682	**47.96/0.789**	0.08/0.442	1.58/0.498	1.96/0.486	**13.20/0.650**	19.03/0.653	13.85/0.596
GDPS-LV2 (ref.)	8.54/0.567	18.27/0.684	47.04/0.786	0.07/0.389	1.82/0.501	1.98/0.486	11.91/0.645	17.35/0.642	13.37/0.587

Table 7: Official IWSLT 2026 blind test-set results for our primary submission (team SLC, Prior-BLEU+Consensus). yor/hau/ibo are scored with sp-BLEU, character-level chrF, and SSA-COMET; gle is scored with BLEU and chrF++. Metric definitions differ from the validation-set tables; ckb is omitted due to a submission-file issue.

Lang	**spBLEU/BLEU**	**chrF/chrF++**	**SSA-COMET**
yor	11.7	39.8	0.552
hau	2.4	24.3	0.285
ibo	3.8	27.6	0.349
gle	2.4	0.1	–

Key observations: GDPS-Med shows largest gains over UFT-Med on bem (+2.07 BLEU), hau (+1.50 BLEU), aeb (+0.72 BLEU), and ibo (+0.38 BLEU), while UFT-Med is stronger on ckb (10.67 vs 8.38), est (40.06 vs 39.09), and yor (8.43 vs 8.03). The ckb result is consistent with its outlier status in gradient analysis (highest self-similarity 0.509, lowest Gini coefficient 0.74), and Prior-BLEU+Consensus reranking recovers ckb performance to 20.26 BLEU (Table 5).

3.3 Gradient Analysis vs. Final Performance

The gradient-conflict analysis (Figure 2) agrees with final outcomes at the coarse level: ckb is flagged as a clear outlier and is indeed the only language where UFT-Med substantially beats GDPS-Med (-2.29 BLEU). Within the Niger-Congo cluster (hau, ibo, yor), however, the alignment is only partial: despite high pairwise gradient similarity, GDPS-Med gains diverge (hau $+1.50$, ibo $+0.38$, yor -0.40 BLEU). This suggests that pairwise similarity captures whether languages point in compatible directions, but not whether each is equally well aligned with the shared subspace. yor recovers strongly under reranking (8.03→13.23 BLEU), indicating its difficulty is better addressed at inference than by parameter sharing. Current grouping criteria reliably localize strong outliers but have limited power for intra-cluster variation; per-language alignment scores are a promising future direction.

3.4 Ablation: 7-Language Robustness

Table 4 compares 8-language vs. 7-language (ckb removed). The effect is *capacity-dependent*: for GDPS-LV2, 6 of 7 languages improve (aeb $+0.85$, bem $+1.30$, est $+0.96$, hau $+0.48$, ibo $+0.33$; only yor -0.34), showing that with sufficient capacity, removing ckb's conflicting gradients benefits the remaining languages. For Medium models and UFT-LV2, changes are mixed and small (± 0.5 BLEU, near noise). We therefore view ckb as a genuine outlier whose removal is at worst harmless and, for higher-capacity models, mildly beneficial.

3.5 Ablation: GDPS Hyperparameters (K and Shared Ratio)

To examine the sensitivity of GDPS to its two key hyperparameters, we conduct a single-factor ablation on the LargeV2 backbone: the number of language clusters K and the shared-private ra-

Table 8: BLEU/chrF++ on 6-language official subset. 6L=trained on the six IWSLT 2026 official directions only; 8L=trained on all 8 directions (including aeb, est). All systems use pre-trained models (unconstrained condition).

	Lower-conflict languages					High-confl.
Setting	**bem**	**gle**	**hau**	**ibo**	**yor**	**ckb**
6L-UFT-Med	16.08/37.57	0.07/9.67	1.03/16.81	2.04/17.23	6.91/26.17	7.81/25.98
6L-GDPS-Med	18.17/41.04	0.06/10.43	1.35/18.66	1.93/17.67	5.60/25.63	7.15/24.84
6L-UFT-LV2 (ref.)	18.03/39.63	0.08/10.53	1.55/19.52	1.87/18.73	**12.36/33.93**	**15.51/37.18**
6L-GDPS-LV2 (ref.)	17.89/38.97	0.10/11.52	1.52/19.34	1.64/17.82	10.91/32.74	15.29/36.57
8L-UFT-Med	18.49/41.58	0.10/10.38	1.80/20.64	3.00/20.44	8.43/29.39	10.67/29.17
8L-GDPS-Med	**20.56/43.70**	**0.16/10.14**	**3.30/23.38**	**3.38/20.76**	8.03/29.20	8.38/26.46

tio ρ (Eq. (4)). The default configuration (K=4, ρ=0.50) is determined by our gradient analysis. Table 9 compares alternatives, holding one parameter fixed at its default while varying the other.

Impact of Cluster Count K: The choice of K has a clear and predictable effect consistent with the gradient analysis. At K=2, the high-conflict outlier **ckb** is merged into the dominant shared cluster, suffering a substantial BLEU drop from 17.35 to 15.51 (-10.6%) due to unresolved gradient interference—precisely the failure mode our clustering is designed to prevent. At K=3, the dendrogram cut falls at a suboptimal boundary that merges incompatible language pairs and disrupts natural cluster structure, causing **gle** to collapse to 7.17 chrF++ (-31% relative to the default). These failure modes directly correspond to the two risks identified in Section 2.3.2: under-clustering exposes outliers to gradient conflict, while an ill-placed cut boundary disrupts positive transfer within compatible groups. The gradient-driven K=4 avoids both.

Impact of Shared-Private Ratio ρ: The sensitivity to ρ is comparatively modest across the $[0.25, 0.75]$ range, with most per-language differences within ± 0.5 BLEU. The clearest signal is directional: reducing ρ to 0.25 provides additional private capacity that marginally benefits **ckb** (17.75 BLEU) but weakens cross-lingual transfer for low-resource directions such as **bem** (16.96 vs. 18.27) and **hau** (1.56 vs. 1.82). Increasing ρ to 0.75 reverses this, recovering **hau** and **ibo** at the cost of **ckb** (16.94 BLEU). The gradient-derived value of ρ=0.50 falls within the stable plateau where neither direction collapses, confirming that the conflict score provides a useful, if approximate, guide for capacity allocation without requiring exhaustive search.

3.6 Ablation: Curriculum Distillation and Test-Time Reranking

Table 5 ablates curriculum stages and reranking strategies.

Curriculum Distillation: Progressive distillation (Distill-only → mix20 → mix35 → mix50) yields consistent gains; mix50 is best overall, with notable improvements on aeb (6.92→7.60), ckb (3.50→4.50), and hau (1.28→1.52). Estonian remains stable and Irish near-zero.

Test-Time Reranking: Prior-BLEU substantially outperforms both baseline and TT-SCR: it recovers ckb to 19.23 BLEU and yor to 13.20 BLEU, demonstrating that leveraging historical BLEU priors effectively addresses gradient conflict at inference. TT-SCR shows mixed results, suggesting the prior-free approach struggles when model consensus does not align with quality. Prior-BLEU+Consensus matches or slightly improves upon Prior-BLEU alone.

3.7 Additional Evaluation: COMET

Table 6 presents COMET scores. COMET correlates closely with human judgment and complements BLEU/chrF++.

The COMET ranking is broadly consistent with BLEU: GDPS-Med attains the highest average COMET (0.573 vs. 0.568), echoing its BLEU advantage, with the strongest agreement on bem and hau. The divergences are more informative: on est and yor, GDPS-Med trails slightly on BLEU yet COMET scores are tied, indicating the BLEU gaps reflect surface-form variation rather than adequacy loss. This suggests GDPS preserves meaning where n-gram overlap dips. The main exception is ckb, where both metrics agree the Medium models lag, consistent with its outlier profile.

Table 9: GDPS hyperparameter ablation on LargeV2 (8-language setup), reported as BLEU/chrF++. The default configuration (K=4, ρ=0.50), automatically determined by our gradient analysis, is marked with †. When varying K, ρ is fixed at 0.50; when varying ρ, K is fixed at 4. Avg = macro-average across all 8 languages.

	Lower-conflict languages							High-confl.	
Configuration (K, ρ)	**aeb**	**bem**	**est**	**gle**	**hau**	**ibo**	**yor**	**ckb**	**Avg**
Default configuration (gradient-driven)									
GDPS-LV2† (4, 0.50)	8.54/26.31	18.27/40.67	47.04/67.29	0.07/10.33	1.82/20.31	1.98/18.81	11.91/34.00	17.35/38.59	13.37/32.04
Varying number of clusters K (with $\rho = 0.50$)									
Varying K (2, 0.50)	9.17/27.07	19.38/40.60	47.53/67.71	0.07/11.40	1.81/20.46	1.91/19.02	11.41/33.50	15.51/37.06	13.35/32.10
Varying K (3, 0.50)	8.87/26.34	16.84/39.14	47.72/67.84	0.09/7.17	1.69/19.75	1.67/18.17	11.35/33.23	17.67/39.20	13.24/31.36
Varying shared-private ratio ρ (with $K = 4$)									
Varying ρ (4, 0.25)	8.57/26.39	16.96/39.42	47.39/67.59	0.08/12.73	1.56/19.51	1.85/18.31	11.34/33.72	17.75/39.32	13.19/32.13
Varying ρ (4, 0.75)	9.04/27.23	18.15/39.23	47.24/67.55	0.08/12.97	2.18/19.67	2.11/18.53	11.87/33.46	16.94/38.58	13.45/32.15

3.8 Official Test-Set Results

Table 7 reports the official IWSLT 2026 blind test-set scores for our primary submission (team SLC, Prior-BLEU+Consensus). The official metrics differ from validation: yor/hau/ibo use spBLEU, character-level chrF, and SSA-COMET; gle uses BLEU and chrF++. Absolute values are not directly comparable to our validation numbers, but the relative picture is consistent: yor reaches 11.7 spBLEU, and hau/ibo remain low as expected. The ckb-eng submission returned an anomalously low score and is omitted; we attribute this to a submission-file formatting issue rather than to the model itself, since the same checkpoints recover strong ckb performance on our validation set (Table 5).

Table 8 presents results on the 6-language official subset, comparing models trained on 6 directions only (6L) versus models trained on all 8 directions (8L).

4 Conclusion

This paper presented the CUHKSZ system for the IWSLT 2026 Low-Resource Speech-to-Text task. Our key contribution is GDPS (Gradient-Driven Parameter Sharing), a framework that applies gradient analysis to automatically identify language groupings and optimal shared-private parameter ratios. Unlike uniform fine-tuning, GDPS achieves substantial gains on languages with strong gradient conflicts: bem (+2.07 BLEU), hau (+1.50 BLEU), and ibo (+0.38 BLEU) compared to UFT-Med. Conversely, languages with isolated gradient profiles (ckb, yor) benefit more from prior-aware test-time reranking at inference.

Our experimental findings highlight four key insights: (1) GDPS specialization recovers performance for conflict-prone languages while maintaining gains via Prior-BLEU+Consensus reranking (ckb: 8.38→20.26, yor: 8.03→13.23); (2) the high-conflict outlier ckb (lowest silhouette and Gini) can be safely excluded: its removal is at worst harmless and, for the higher-capacity GDPS-LV2 model, improves 6 of the 7 remaining languages, though the effect is small and mixed for the lower-capacity settings; (3) expanded language coverage (8L vs. 6L training) amplifies GDPS gains, particularly for bem, hau, and ibo; and (4) curriculum distillation and test-time reranking strategies provide complementary improvements, with Prior-BLEU+Consensus achieving 14.85 average BLEU.

While our gradient analysis reliably localizes strong outliers such as ckb, it is less predictive of intra-cluster variation: GDPS gains range from $+1.50$ to -0.40 BLEU within the Niger-Congo group alone.

Acknowledgments

This paper is supported by Project W2531054 of the National Natural Science Foundation of China, and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, and 1 others. 2023. Findings of

the IWSLT 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, and 1 others. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, and 1 others. 2023a. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, and 1 others. 2023b. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351.

Xin Dong, Ruize Wu, Chao Xiong, Hai Li, Lei Cheng, Yong He, and 1 others. 2022. Gdod: Effective gradient descent using orthogonal decomposition for multi-task learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 386–395.

Isaac Gerber. 2025. Attention is not all you need: The importance of feedforward networks in transformer models. *arXiv preprint arXiv:2505.06633*.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Chutong Meng and Antonios Anastasopoulos. 2025. GMU Systems for the IWSLT 2025 Low-Resource Speech Translation Shared Task. In *Proc. IWSLT*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25:1–52.

Nathaniel R. Robinson, Niyati Bafna, Xiluo He, Tom Lupicki, Lavanya Shankar, Cihan Xiao, Qi Sun, Kenton Murray, and David Yarowsky. 2025. JHU IWSLT 2025 Low-resource System Description. In *Proc. IWSLT*.

SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637:587–593.

Claytone Sikasote, Eunice Mukonde, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. BIGC: A Multimodal Multi-Purpose Dataset for Bemba. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2062–2078.

Tiia Sildam, Andra Velve, and Tanel Alumäe. 2024. Finetuning end-to-end models for Estonian conversational spoken language translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 166–174.

Ruiyan Sun and Satoshi Nakamura. 2026. Gradient-informed training for low-resource multilingual speech translation. *arXiv preprint arXiv:2603.25836*.

Ziyu Yang, Guibin Chen, Yuxin Yang, Aoxiong Zeng, and Xiangquan Yang. 2026. Disentangling task conflicts in multi-task lora via orthogonal gradient projection. *arXiv preprint arXiv:2601.09684*.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836.

Shijie Zhu, Hui Zhao, Tianshu Wu, Pengjie Wang, Hongbo Deng, Jian Xu, and Bo Zheng. 2025. Gradient deconfliction via orthogonal projections onto subspaces for multi-task learning. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*.