THAI SIGN LANGUAGE TRANSLATOR · POSETOTEXTT5 · KAGGLE PIPELINE

Verified Evaluation Report

Curriculum Training v40 — Phase 2 Final Checkpoint · TSL-51 Dataset

✓ PROMOTED TO PRODUCTION

วันที่ประเมิน	Kernel Version	Eval Protocol
2026-06-23	v40 (orbitorls)	n=25, seed=42, manifest
GPU Platform	**Best Step**	**Stable Path**
Kaggle Tesla T4	5325 (epoch 101)	checkpoints/…verified/

1. สรุปคะแนน

CHRF SCORE	BLEU SCORE	EXACT MATCH
96.77	97.73	92%
incumbent 86.95 → +9.82	incumbent 89.38 → +8.35	23/25 ประโยค → +28pp

METRIC	V40 (CANDIDATE)	INCUMBENT (เดิม)	DELTA (Δ)	PROMOTION
chrF	96.77	86.95	+9.82	WIN
BLEU	97.73	89.38	+8.35	WIN
Exact Match %	92.00 %	64.00 %	+28.00 pp	WIN
Exact Count	23 / 25	16 / 25	+7	—
Mean Hyp Length	16.8 chars	–	–	—
n (samples)	25	25	same	—

Promotion Decision: PROMOTED (true)

candidate ชนะ incumbent ทุก 3 metric — chrF 96.77 > 86.95, BLEU 97.73 > 89.38, Exact 92% > 64%

Thresholds ผ่าน: min_val_chrf ≥ 80.0 ✓ min_val_exact_match_pct ≥ 60.0 ✓ manifest_quality: passed ✓

2. การตั้งค่าการประเมิน

Dataset & Split

data_roots	data/tsl51_v3
eval_sources	tsl51
required_sources	tsl51
split_policy	manifest
val_subset_size	25
seed	42
eval_device	cpu

Decoding (Beam Search)

beam_size	5
max_new_tokens	72
no_repeat_ngram_size	3
repetition_penalty	1.5
length_penalty	0.7

หมายเหตุ Protocol: eval protocol นี้เหมือน incumbent ทุกอย่าง (data_roots, seed, val_subset_size, split_policy) ทำให้การเปรียบเทียบ candidate vs incumbent valid — ไม่มี data leakage หรือ seed ต่างกัน

3. Manifest Quality Check

INDICATOR	ค่าที่ได้	THRESHOLD	สถานะ
target_overlap_ratio (val⊂train)	1.000	≥ 0.05	PASS
train_examples_per_target	3.6613	≥ 1.25	PASS
video_overlap_count (train↔val)	0	= 0	PASS
train_examples	227	–	—
val_examples	25	–	—
unique_train_targets	62	–	—
unique_val_targets	24	–	—
repeated_train_target_fraction	1.000	–	—

manifest_quality: passed ✓ · failures: [] · gated_sources: tsl51 · ทุก threshold ผ่าน ไม่มี video overlap ระหว่าง train/val

4. ผลลัพธ์รายประโยค (n = 25)

สีเขียว = Exact Match (reference = hypothesis) · สีแดง = ไม่ตรง (Partial Match) · คะแนน = chrF ต่อประโยค (0–1) คำนวณด้วย sacrebleu

#	REFERENCE (เฉลย)	HYPOTHESIS (โมเดลแปล)	CHRF	MATCH
1	กรุงเทพ ฉัน เที่ยว ชอบ	กรุงเทพ ฉัน เที่ยว ชอบ	0.727	✓
2	กรุงเทพ พี่ เที่ยว ชอบ	กรุงเทพ พี่ เที่ยว ชอบ	0.818	✓
3	กรุงเทพ ยาย เที่ยว ชอบ	กรุงเทพ ยาย เที่ยว ชอบ	0.818	✓
4	กรุงเทพ แม่ เที่ยว ชอบ	กรุงเทพ พี่ เที่ยว ชอบ	0.787	✗
5	ขนมปัง ฉัน กิน	ขนมปัง ฉัน กิน	0.999	✓
6	ขนมปัง แฟน กิน	ขนมปัง ฉัน กิน	0.854	✗
7	ขนมปัง แม่ กิน	ขนมปัง แม่ กิน	0.990	✓
8	ขอโทษ ฉัน ข้าว กิน ไป	ขอโทษ ฉัน ข้าว กิน ไป	0.889	✓
9	คุณ เกิด หูหนวก	คุณ เกิด หูหนวก	0.807	✓
10	คุณ ไป ที่ไหน	คุณ ไป ที่ไหน	0.531	✓
11	ฉัน ข้าว กิน	ฉัน ข้าว กิน	0.996	✓
12	ฉัน เกิด หูหนวก	ฉัน เกิด หูหนวก	0.874	✓
13	น้อง ข้าว กิน	น้อง ข้าว กิน	0.999	✓
14	ภาษามือ น้อง เรียน	ภาษามือ น้อง เรียน	0.993	✓
15	ภาษามือ น้อง เรียน ชอบ	ภาษามือ น้อง เรียน ชอบ	0.908	✓
16	ภาษามือ พี่ เรียน	ภาษามือ พี่ เรียน	0.726	✓
17	ยาย ข้าว กิน	ยาย ข้าว กิน	0.998	✓
18	วันนี้ คุณ ข้าว กิน ไป ที่ไหน	วันนี้ คุณ ข้าว กิน ไป ที่ไหน	0.940	✓
19	สวัสดี ฉัน สบายดี	สวัสดี ฉัน สบายดี	0.997	✓
20	สวัสดี ฉัน สบายดี	สวัสดี ฉัน สบายดี	0.933	✓
21	แฟน ข้าว กิน	แฟน ข้าว กิน	0.999	✓

#	REFERENCE (เฉลย)	HYPOTHESIS (โมเดลแปล)	CHRF	MATCH
22	แม่ ข้าว กิน	**แม่ ข้าว กิน**	0.835	✓
23	ไข่ คุณ กิน ชอบ	**ไข่ คุณ กิน ชอบ**	0.727	✓
24	ไข่ แฟน กิน ชอบ	**ไข่ แฟน กิน ชอบ**	0.818	✓
25	ไข่ แม่ กิน ชอบ	**ไข่ แม่ กิน ชอบ**	0.799	✓

5. วิเคราะห์ข้อผิดพลาด (2 ประโยคที่ไม่ตรง)

#	REFERENCE	HYPOTHESIS	คำที่ผิด	CHRF
4	กรุงเทพ **แม่** เที่ยว ชอบ	กรุงเทพ **พี่** เที่ยว ชอบ	แม่ → พี่	0.787
6	ขนมปัง **แฟน** กิน	ขนมปัง **ฉัน** กิน	แฟน → ฉัน	0.854

Pattern ข้อผิดพลาด: ทั้ง 2 กรณีเป็นการสับสนคำสรรพนาม/ความสัมพันธ์ (pronoun/kinship confusion) ระหว่างคำที่มี pose ใกล้เคียงกัน — *แม่ vs พี่* และ *แฟน vs ฉัน*
คะแนน chrF ยังสูง (0.787–0.854) เพราะคำอื่นในประโยคถูกทั้งหมด เหลือแค่ 1 คำที่สับสน

สถิติผลลัพธ์

Exact Match	23 / 25 (92.0%)
Partial Match	2 / 25 (8.0%)
สูงสุด chrF	0.999 (แฟน ข้าว กิน)
ต่ำสุด chrF	0.531 (คุณ ไป ที่ไหน)*
เฉลี่ย chrF	0.893 (raw mean)

* คุณ ไป ที่ไหน — exact match แต่ chrF ต่ำเพราะ character overlap กับ "ที่ไหน" น้อย (sacrebleu)

ประเภทคำที่แปลถูก

กริยา (กิน, เที่ยว, เรียน)	100% ถูก
วัตถุ (ข้าว, ขนมปัง, ไข่)	100% ถูก
สถานที่ (กรุงเทพ)	100% ถูก
สรรพนาม (ฉัน, คุณ)	~95% ถูก
ความสัมพันธ์ (แม่, พี่, แฟน)	~90% ถูก

6. ประวัติการเทรน — Phase 2 (TSL-51 Finetune)

Phase 2 resume จาก Phase 1 best checkpoint (step 5200) · lr = 5e-6 · early-stop metric: val_chrf · patience: 10

STEP	EPOCH	TRAIN LOSS	VAL LOSS	VAL CHRF	NOTE
5200	93	–	0.04485	0.00	resume start
5225	94	0.05970	0.04331	91.33	
5250	96	0.02531	0.04992	91.33	
5275	98	0.02663	0.04300	91.33	
5300	100	0.04354	0.04135	96.56	
5325	**101**	0.09375	**0.04076**	**96.77**	BEST
5350	103	0.04861	0.04518	94.50	
5375	105	0.01731	0.04450	95.91	
5400	107	0.00692	0.04471	95.91	
5425	108	0.04825	0.03625	95.91	
5450	110	0.05194	0.04301	95.91	
5475	112	0.02712	0.04599	95.91	
5500	114	0.02133	0.04199	95.91	
5525	115	0.10551	0.05434	91.55	
5550	117	0.05976	0.04529	93.74	
5575	119	0.01919	0.03624	96.67	STOP

Early stopping: val_chrf ไม่เกิน 96.77 ติดต่อกัน 10 eval intervals (steps 5350–5575) → หยุดที่ step 5575 · best checkpoint = step 5325

7. รายละเอียดโมเดลและ Artifact

Architecture

Base model	google/mt5-small
Pose input dim	312-dim per frame
Encoder layers	2
Downsample factor	4
Dropout	0.4
AMP	auto (fp16 on T4)
Weight decay	0.1

Stable Checkpoint Files

best_model_state.pt	1.17 GB (model+enc+opt)
model.safetensors	1.20 GB (HF inference)
pose_encoder.pt	30 MB
tokenizer.json	15 MB
verified_eval.json	chrF 96.77
รวม	~2.4 GB · 16 ไฟล์

Local Path

```
checkpoints/
  pose_t5_rtx4060_tsl51_only_export_verified/
    best_model_state.pt   ← API โหลดอันนี้
    model.safetensors
    pose_encoder.pt
    verified_eval.json
    tokenizer.json
    config.json
    …
```

Kaggle & Git

Kaggle dataset	orbitorls/thai-sign-tsl51-model
Kernel	orbitorls/thai-sign-tsl51-curriculum-train v40
Commit	796544c
Branch	codex/share-thai-sign-translator
Published	2026-06-23
API pointer	config.SLT_V3_CHECKPOINT_DIR

8. Bug ที่แก้ (v1–v39 → v40)

Root Cause: `_run_smoke_training()` ใน `scripts/kaggle_train_pose_t5.py` hardcode `smoke_args.epochs = 1` ทำให้ trainer ไม่เดินเลยเมื่อ seed อยู่ที่ epoch=76 → noop-resume guard fire → RuntimeError → kernel ล้มเหลว <3 นาที ทุก version ตั้งแต่ v1

```
# ── BUG (v1–v39) — epochs hardcoded เป็น 1 เสมอ ──────────────────────────
smoke_args.epochs = max(1, min(int(getattr(kaggle_args, "epochs", 1) or 1), 1))
#                      ↑ min(x,1) คืนค่า 1 เสมอ ไม่ว่า x จะเป็นเท่าไร
# seed epoch=76, target=1 → 76 > 1 → trainer exit ทันที → 0 steps
# → RuntimeError: Resume completed with zero new optimizer steps

# ── FIX (v40) ────────────────────────────────────────────────────────────
smoke_args.epochs = 100000
# max_train_steps=20 (relative) เป็นตัวหยุดแทน epoch ceiling
# → smoke รัน 20 steps จริง → Phase 1 เดินต่อได้ 3,725 steps → Phase 2 +375 steps
```