SELECTIVE TARGET GEOMETRY FOR TEST-TIME MULTIMODAL SENTIMENT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting a multimodal sentiment model to an unlabeled target corpus is difficult when language, recording conditions, and annotation conventions change together. Source-free methods that align target features with a reconstructed source distribution may reinforce a mismatched geometry, while target-only pseudo-labeling can corrupt prototypes when initial predictions are unreliable. We present *Selective Target Geometry* (STG), an offline test-time adaptation method that initializes class-conditional prototypes from statistics saved with the source model and selectively updates them using unlabeled target batches. A frozen text teacher supplies semantic anchors, lightweight audio-visual evidence moderates uncertain pseudo-labels, and a router reuses, creates, or defers prototype updates according to batch compatibility. We evaluate ten directed transfers among four English and Chinese multimodal sentiment datasets. In the reported results, STG reduces mean absolute error relative to the frozen source model in all ten transfers and achieves the lowest mean absolute error among the compared methods in each transfer. An ablation on MOSI-to-SIMS shows that removing the router largely eliminates the gain. These findings suggest that selective updates to target geometry can reduce negative transfer when source-alignment assumptions are unreliable.

1 INTRODUCTION

Multimodal sentiment models combine text, voice, and facial cues, but transfer poorly when a target corpus changes language, recording conditions, or annotation conventions. We study offline test-time domain adaptation: the unlabeled target training split may be traversed repeatedly without access to source examples, and a disjoint target test split is reserved for evaluation (Liang et al., 2025).

Source-free methods can reconstruct a source-like distribution and align target features to it (Tian et al., 2021; Qiu et al., 2021). That reference may be poorly matched under cross-lingual transfer; abandoning it entirely leaves target prototypes dependent on uncertain pseudo-labels. The question is *when to retain source structure, adapt it, or withhold an update*.

We propose *Selective Target Geometry* (STG). Class-conditional source statistics initialize a library of prototype sets; a frozen text pathway supplies semantic labels and grouped acoustic and facial descriptors provide weak evidence. A router then decides whether each batch should reuse a prototype set, create one, or defer updating. Source statistics are an initialization, not a fixed alignment target; no target labels enter adaptation.

Across ten directed transfers among MOSI, MOSEI, CH-SIMS, and CH-SIMSv2, the reported mean absolute error (MAE) is lower than for the unadapted model and five adaptation baselines in every scenario. Removing the router causes the largest degradation in the MOSI-to-SIMS ablation. We distinguish cross-lingual from same-language transfer by provenance, without treating language direction as a measured gap magnitude.

2 RELATED WORK

Multimodal sentiment analysis. Fusion models have progressed from tensor fusion (Zadeh et al., 2017) to cross-modal transformers (Tsai et al., 2019). MOSI and MOSEI provide English video

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

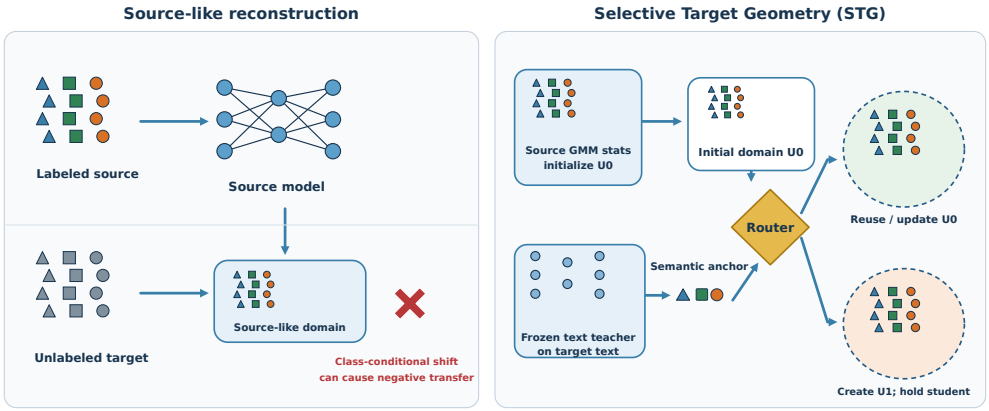

Figure 1: **STG paradigm overview. (Left)** Some source-estimation methods construct a source-like virtual distribution and force target features to align with it—an approach that can fail when $P_s(Z \mid Y) \not\approx P_t(Z \mid Y)$ in cross-lingual scenarios. **(Right)** STG initializes geometry from saved source statistics and selectively adapts it using unlabeled target batches. A frozen text teacher provides semantic labels; the router reuses, creates, or defers prototype updates.

sentiment data (Zadeh et al., 2016; Bagher Zadeh et al., 2018), while CH-SIMS and CH-SIMSv2 provide Chinese data with modality-specific annotations (Yu et al., 2020; Liu et al., 2022). STG keeps a late-fusion backbone and addresses adaptation across these corpora, rather than proposing another fusion architecture.

Unlabeled test-time adaptation. The test-time domain adaptation (TTDA) setting permits multiple passes over an unlabeled target set, unlike single-pass online adaptation (Liang et al., 2025). Parameter-update methods such as Tent minimize prediction entropy (Wang et al., 2021); EATA and SAR seek more reliable updates under uncertain predictions (Niu et al., 2022; 2023). SHOT instead transfers the source classifier and exploits target pseudo-labels (Liang et al., 2021). These approaches motivate our concern with feedback from erroneous target predictions, but do not specify when a multimodal prototype set should be reused, expanded, or left unchanged.

Target-side structure without source examples. When source data are unavailable, a complementary strategy extracts structure from target features. T3A adjusts a classifier using target pseudo-prototypes without backpropagation (Iwasawa & Matsuo, 2021). Neighborhood methods use local and reciprocal neighbors to encourage target label consistency (Yang et al., 2021), while TAST combines neighbor-based pseudo-labels with prototype distributions (Jang et al., 2023). Divide and Contrast distinguishes relatively source-like, confident target samples from target-specific ones, using different global and local objectives (Zhang et al., 2022). This line makes the trade-off between global pseudo-label errors and uncertain local relations explicit. STG does not propose a new neighborhood objective; its router instead determines whether a batch changes an existing geometric reference, warrants a separate one, or should leave the student unchanged. These classification-oriented methods also do not directly settle how to represent continuous sentiment under joint text, audio, and visual shift.

Source knowledge and multimodal reliability. VDM-DA approximates unavailable source features with a virtual distribution, and CPGA generates avatar prototypes for source-free adaptation (Tian et al., 2021; Qiu et al., 2021). STG likewise retains source-derived statistics in the checkpoint, but uses them to initialize target-side geometry instead of fixing a source-alignment objective. READ studies modality reliability bias in multimodal TTA (Yang et al., 2024). Closer to sentiment regression, CASP combines contrastive adaptation with stable pseudo-labeling (Guo et al., 2025), and GMEL models sentiment groups with a von Mises–Fisher mixture and multiscale re-dropout (Tang et al., 2026). GMEL's group structure and augmentation seek more robust sentiment representations; STG instead adapts a class-conditional Gaussian library initialized from shipped source statistics. Its batch-level REUSE, CREATE, and DEFER decisions govern whether that geometry and

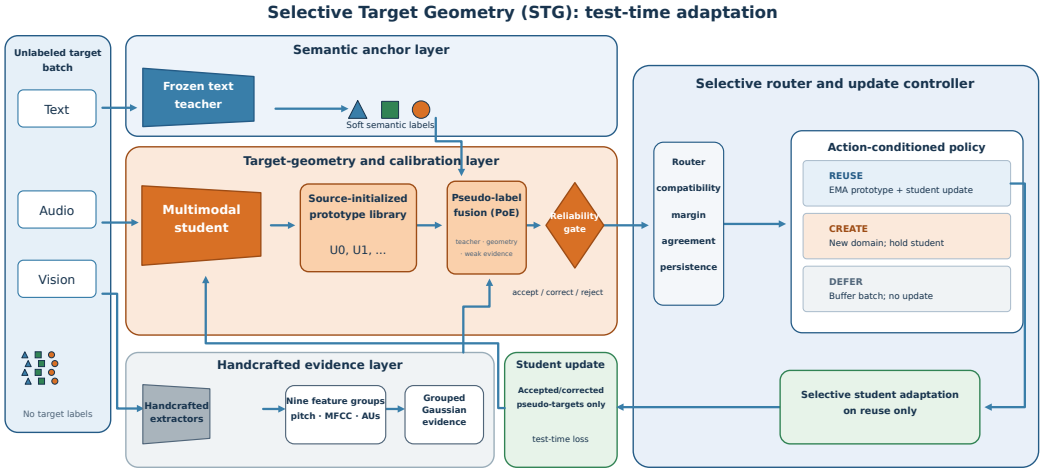

Figure 2: **STG framework overview.** A frozen text teacher supplies soft class labels; a source-initialized Gaussian library adapts toward target features; nine acoustic–visual feature groups provide weak evidence for pseudo-label reliability. The router reuses an existing domain, creates a new one, or defers the batch. Only reuse permits a student update; creation holds the student fixed.

the student are updated. Acoustic–visual descriptors serve only as weak evidence in those decisions. This identifies a different adaptation mechanism, not a claim that mixture modeling or target structure is new. The reported comparisons test the resulting design against these complementary approaches.

3 Selective Target Geometry

Setting. Let a source-trained multimodal regressor be $F_s = C_s \circ G_s(E_T, E_A, E_V)$, with fusion feature $h \in \mathbb{R}^D$. Source training uses labeled text–audio–vision examples; adaptation has only the checkpoint, statistics saved with it, and an unlabeled target corpus. We use the offline test-time domain adaptation setting: target training batches may be revisited, but target test labels never enter adaptation. A student $(\hat{F}, \hat{C})$ is initialized from the source model. The class-coordinate system uses $K = 3$ sentiment centers $c_k \in \{-1, 0, +1\}$.

3.1 A Fixed Semantic Anchor and a Moving Geometric Reference

We restrict the frozen source model to its text pathway and average R Monte Carlo dropout predictions. We then map the continuous output to soft class coordinates:

$$p_i^T = \frac{1}{R}\sum_{r=1}^{R} F_s^T(x_i^T; \text{dropout}^{(r)}), \qquad q_i^T(k) = \frac{\exp[-\gamma(p_i^T - c_k)^2]}{\sum_{l=1}^{K}\exp[-\gamma(p_i^T - c_l)^2]}. \tag{1}$$

The text pathway is an anchor, not a second student: it stays frozen as the fusion model adapts. Its confidence is $w_i^T = \max_k q_i^T(k)$.

During source training, labeled fusion features initialize a class-conditional Gaussian library $\mathcal{V} = \{\mathcal{U}_e\}_{e=0}^{N_v}$, with each $\mathcal{U}_e = \{(\mu_{e,k}, \Sigma_{e,k}, \pi_{e,k})\}_{k=1}^{K}$. For the source domain,

$$\mu_{0,k} = \frac{\sum_{i\in\mathcal{S}} \mathbf{1}[y_i \mapsto k] h_i}{\sum_{i\in\mathcal{S}} \mathbf{1}[y_i \mapsto k] + \epsilon}, \qquad \pi_{0,k} = \frac{\sum_{i\in\mathcal{S}} \mathbf{1}[y_i \mapsto k]}{N_s}. \tag{2}$$

The diagonal covariance uses shrinkage toward a spherical prior. The resulting $\mathcal{U}_0$ ships with the checkpoint. At adaptation time the posterior of feature h_i under sub-domain e is

$$q_{i,e}^V(k) = \frac{\pi_{e,k}\exp[-d_{e,k}(h_i)/\kappa]}{\sum_l \pi_{e,l}\exp[-d_{e,l}(h_i)/\kappa]}, \quad d_{e,k}(h) = (h - \mu_{e,k})^\top \Sigma_{e,k}^{-1}(h - \mu_{e,k}), \tag{3}$$

with $\kappa = D$. The chosen domain $e^\star$ supplies $q_i^V = q_{i,e^\star}^V$ and geometric support $s_i^V = \exp[-\min_k d_{e^\star,k}(h_i)/\kappa]$. Reliability-weighted batch estimates update prototype means and priors by

$$\hat{\mu}_{e,k}^B = \frac{\sum_{i\in B} a_i \bar{q}_i(k) h_i}{\sum_{i\in B} a_i \bar{q}_i(k) + \epsilon}, \quad \mu_{e,k} \leftarrow (1-\beta_k)\mu_{e,k} + \beta_k \hat{\mu}_{e,k}^B, \tag{4}$$

where $\beta_k = \beta_0 \,\mathrm{clip}(n_k^{\mathrm{eff}}/n_0, 0, 1)$ and $n_k^{\mathrm{eff}} = \sum_{i\in B} a_i \bar{q}_i(k)$. Priors and diagonal covariances follow analogous EMA updates; the full definitions are in Appendix A. The source distribution is therefore an initialization, not a fixed alignment target.

3.2 Weak Audio-Visual Evidence and Pseudo-Labels

Ninety-five handcrafted audio-visual descriptors are partitioned into nine semantic groups. Each group fits a diagonal Gaussian $P_{\phi_g}(r_i^{(g)} \mid Y = k)$ on a warm-up segment of the *unlabeled* target corpus using text-teacher soft labels. The descriptors never enter the student's forward pass. Their log evidence and agreement-based reliability are

$$E_{i,g,k} = \log \frac{\pi_k P_{\phi_g}(r_i^{(g)} \mid Y = k)}{\sum_{l\neq k} \pi_l P_{\phi_g}(r_i^{(g)} \mid Y = l)}, \quad \omega_g \leftarrow \eta\omega_g + (1-\eta)\mathbb{E}_{i\in B}[\mathbf{1}(\arg\max_k E_{i,g,k} = \hat{k}_i)], \tag{5}$$

where $\hat{k}_i$ is the student's binned class. The weighted evidence is $E_i^{\mathrm{hand}}(k) = \sum_g \omega_g E_{i,g,k}$. Evidence consensus yields $\mathrm{cons}_i \in [0,1]$ and batch mean $\bar{c}(B)$; its exact definition and the sample accept/correct/reject cases are in the appendix.

Text, geometry, and evidence are fused through a weighted product of experts:

$$\log \bar{q}_i(k) \propto w_i^T \log q_i^T(k) + s_i^V \log q_i^V(k) + \lambda_H E_i^{\mathrm{hand}}(k), \qquad \bar{y}_i = \sum_{k=1}^K c_k \bar{q}_i(k). \tag{6}$$

The use of the centers c_k, rather than the class indices k, preserves the continuous $[-1,1]$ sentiment scale. The sample weight is $a_i = w_i^T s_i^V [1 - \overline{\mathrm{JS}}(q_i^T \| q_i^V)](1 - \mathrm{conflict}_i)$, where $\overline{\mathrm{JS}}$ is normalized Jensen–Shannon divergence.

3.3 Selective Routing and Student Updates

A batch is scored against each prototype set by teacher-weighted distance and class separation:

$$C_e(B) = -\frac{1}{|B|}\sum_{i\in B}\sum_k q_i^T(k)\frac{d_{e,k}(h_i)}{\kappa} + \lambda_{\mathrm{sep}} \min_{j\neq l} \|\mu_{e,j} - \mu_{e,l}\|_2. \tag{7}$$

Let $e^\star = \arg\max_e C_e(B)$, $\bar{C}(B) = C_{e^\star}(B)$, and ΔC be the margin over the second-best domain when one exists. The router chooses

$$r(B) = \begin{cases} \textsc{defer}, & \bar{c}(B) < \tau_{\mathrm{cons}}, \\ \textsc{reuse}(e^\star), & C_{e^\star}(B) > \tau_{\mathrm{reuse}} \text{ and } \Delta C > \delta, \\ \textsc{create}, & \bar{C}(B) < \tau_{\mathrm{create}} \text{ for } L \text{ batches}, \\ \textsc{defer}, & \text{otherwise}. \end{cases} \tag{8}$$

For the initial single-domain case, the reuse-margin condition requires an explicit convention in the implementation; this is one of the configuration details that must be verified from the run record. Reuse applies prototype EMA and a reduced-rate student update. Create fits a new target subdomain from buffered samples but holds the student fixed. Defer buffers the batch without an update. Domain capacity, cooldown, merge, and pruning rules appear in Appendix A.

Only accepted or corrected samples supervise the student's normalization parameters. The objective is

$$\mathcal{L}_{\mathrm{TTA}} = \mathcal{L}_{\mathrm{pseudo}} + \lambda_1 \mathcal{L}_{\mathrm{proto}} + \lambda_2 \mathcal{L}_{\mathrm{compact}} + \lambda_3 \mathcal{L}_{\mathrm{cons}}, \tag{9}$$

combining regression to $\bar{y}_i$, prototype posterior alignment, feature compactness, and dropout consistency. The appendix provides each term and the full algorithm. This is offline, multi-epoch adaptation; the method does not claim a single-pass online protocol.

Table 1: **Results on all ten adaptation scenarios.** Columns are the ten source→target pairs of Sec. *Adaptation Scenarios*: S1 = MOSI→SIMS, S2 = MOSEI→SIMS, S3 = MOSI→MOSEI, S4 = SIMS→MOSI, S5 = SIMS→MOSEI, S6 = SIMS→SIMSv2, S7 = MOSI→SIMSv2, S8 = MOSEI→SIMSv2, S9 = SIMSv2→MOSI, S10 = SIMSv2→MOSEI. ↓: lower is better, ↑: higher is better; best per column in **bold**.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
MAE ↓										
Source Only	0.6872	0.6991	0.3291	0.6870	0.5677	0.4582	0.5419	0.5224	0.4075	0.3926
Reliable Fusion	0.6867	0.6067	0.3394	0.6176	0.4867	0.4503	0.5363	0.4946	0.3573	0.3458
CASP	0.5788	0.6628	0.3200	0.6547	0.5325	0.4256	0.4985	0.4860	0.3105	0.3280
VDM-DA	0.5820	0.6137	0.3045	0.5828	0.5216	0.4335	0.4964	0.5118	0.3922	0.3841
CPGA	0.6046	0.6329	0.3008	0.5131	0.4796	0.4350	0.4877	0.4712	0.3136	0.2927
GMEL	0.5531	0.5901	0.2842	0.4625	0.3958	0.4204	0.4715	0.4606	0.3971	0.2823
STG (Ours)	**0.5276**	**0.5606**	**0.2437**	**0.4217**	**0.3720**	**0.4149**	**0.4465**	**0.4542**	**0.3073**	**0.2662**
Acc-2 ↑										
Source Only	51.52	56.46	67.37	47.52	47.06	66.08	64.99	54.40	77.84	65.53
Reliable Fusion	50.18	61.17	68.52	47.31	48.63	68.92	66.82	56.67	78.29	66.54
CASP	51.21	64.96	69.07	51.63	50.10	68.13	67.31	60.28	77.95	69.72
VDM-DA	54.88	57.89	67.44	48.97	49.01	67.27	65.79	54.58	78.02	68.24
CPGA	53.57	60.62	68.28	53.16	51.22	68.51	69.60	56.27	78.34	70.30
GMEL	56.03	65.14	70.20	55.68	50.36	69.04	71.93	63.38	77.91	71.01
STG (Ours)	**58.15**	**66.80**	**70.41**	**59.89**	**51.47**	**70.21**	**72.53**	**67.13**	**79.48**	**71.97**
F1 ↑										
Source Only	51.95	57.56	67.34	48.17	48.37	67.56	62.14	52.19	77.89	66.89
Reliable Fusion	48.19	59.29	68.33	48.06	49.96	68.12	63.39	54.83	78.64	67.84
CASP	54.01	**67.05**	68.98	50.06	59.60	69.91	69.96	61.48	79.33	70.23
VDM-DA	52.14	59.87	67.63	48.20	52.71	67.99	64.18	53.39	79.17	69.51
CPGA	56.27	61.53	68.21	49.34	49.59	68.47	66.41	62.66	78.32	69.40
GMEL	58.46	64.75	69.74	52.28	57.65	69.81	70.62	64.13	**80.08**	70.38
STG (Ours)	**60.01**	65.48	**70.55**	**55.73**	**60.32**	**70.25**	**72.29**	**65.26**	79.50	**71.30**
Acc@0.5 ↑										
Source Only	37.64	41.14	73.72	47.64	47.62	61.82	47.87	51.62	80.85	65.36
Reliable Fusion	38.54	39.94	70.19	48.65	50.12	62.79	44.10	51.68	81.90	73.09
CASP	32.85	43.86	71.31	48.87	49.39	62.38	55.26	54.40	81.35	80.16
VDM-DA	38.37	41.99	81.18	52.15	57.79	64.32	50.84	55.93	81.38	75.74
CPGA	39.56	43.12	80.94	55.21	52.75	63.16	54.41	51.78	80.65	65.28
GMEL	40.90	46.36	86.01	53.88	64.67	66.27	60.53	57.73	82.22	80.43
STG (Ours)	**42.74**	**50.68**	**89.76**	**58.05**	**70.03**	**67.99**	**62.10**	**58.37**	**82.36**	**82.91**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We use CMU-MOSI (Zadeh et al., 2016) (1,284/229/686 train/val/test), CMU-MOSEI (Bagher Zadeh et al., 2018) (16,326/1,871/4,659), CH-SIMS (Yu et al., 2020) (1,368/456/457), and CH-SIMSv2 (Liu et al., 2022) (2,722/647/1,034). English scores in $[-3, 3]$ are rescaled to $[-1, 1]$; Chinese scores are already in $[-1, 1]$.

Adaptation Scenarios. We report ten directed transfers: two within a language (MOSI→MOSEI and SIMS→SIMSv2) and eight between languages. MOSEI→MOSI and SIMSv2→SIMS were omitted in the original transfer matrix because of dataset-family overlap. The reported opposite directions also require an original clip-ID audit before they can be interpreted as clip-disjoint evaluations.

Protocol and implementation. Following offline TTDA (Liang et al., 2025), adaptation makes 30 passes over the unlabeled target *train* split; target *test* labels are used only for evaluation. The

Table 2: **Ablation study on MOSI→SIMS (cross-lingual).** Each component is independently removed from the full model; the three indented rows isolate the router's individual actions and the choice of a static versus target-adapted source geometry.

Configuration	MAE↓	Acc-2↑	F1↑	A@0.5↑
Full STG	**0.528**	**58.2**	**60.0**	**42.7**
− Routing	0.691	53.2	54.7	39.0
− Create only (single-domain EMA)	0.611	52.6	54.5	40.1
− Defer only	0.629	52.4	53.8	38.3
Fixed source VD (no EMA, no routing)	0.596	53.7	52.2	38.9
− Handcrafted evid.	0.542	55.4	58.5	42.5
− Source VD init.	0.627	52.7	55.4	39.2
Source-only	0.687	51.5	52.0	37.6

same target order and fixed selection recipe are used for STG and the comparators. The backbone projects features to $D = 64$, followed by two-layer unimodal transformers, three-layer cross-modal fusion, and a three-layer regression head. Source pretraining uses MSE, Adam (10^{-3}), batch size 64, and at most 50 epochs; the unimodal encoders are frozen. A class-conditional diagonal Gaussian reference with covariance shrinkage $\lambda = 0.1$ is fitted on source training features and saved with the checkpoint. STG additionally extracts 51 acoustic and 44 facial descriptors in nine groups; these do not enter the backbone forward pass.

Metrics and comparisons. We report MAE on the continuous $[-1, 1]$ scale, Acc-2 and F1 after thresholding predictions and labels at zero, and Acc@0.5 for errors no larger than 0.5. Three internal prototype centers $\{-1, 0, +1\}$ do not discretize the reported regression predictions. The comparisons are Source-Only, the multimodal adaptation methods Reliable Fusion (Yang et al., 2024), CASP (Guo et al., 2025), and GMEL (Tang et al., 2026), plus the source-estimation methods VDM-DA (Tian et al., 2021) and CPGA (Qiu et al., 2021).

4.2 Main Results

Table 1 reports all four metrics for ten transfers; implementation details are in the appendix. STG has the lowest reported MAE and highest Acc-2 and Acc@0.5 in all ten, and the highest F1 in eight. The F1 exceptions are MOSEI→SIMS (CASP 67.05; STG 65.48) and SIMSv2→MOSI (GMEL 80.08; STG 79.50).

Relative to Source-Only, MAE falls from 0.329 to 0.244 on the within-language MOSI→MOSEI transfer and from 0.458 to 0.415 on SIMS→SIMSv2. The cross-language MOSI→SIMS and SIMS→MOSI transfers show larger absolute reductions, 0.687→0.528 and 0.687→0.422, respectively. These numbers describe the observed transfers; language direction alone does not establish the magnitude or cause of their shift.

4.3 Ablation Study

Table 2 and Figure 3 report component ablations on MOSI→SIMS across all four metrics.

Removing the router raises MAE from 0.528 to 0.691, near the Source-Only value of 0.687. Disabling CREATE alone raises it to 0.611; disabling DEFER raises it to 0.629; fixing the source geometry at its initial state gives 0.596. The effects are not additive, so these ablations support the importance of selective updates without assigning a separate causal share of the total gain to each action.

Replacing source initialization with target pseudo-label initialization raises MAE to 0.627. Removing handcrafted evidence raises it to 0.542, a smaller change; corresponding Acc-2, F1, and Acc@0.5 values appear in the table. The ablations are consistent with a useful source reference



Figure 3: **Ablation study on MOSI→SIMS (cross-lingual).** Panels show **(a)** MAE, **(b)** Acc-2, **(c)** F1, and **(d)** Acc@0.5. The dashed vertical guide marks the source-only result in each panel. Removing routing increases MAE from 0.528 to 0.691; the source-only MAE is 0.687.

and a weak moderating role for the auxiliary descriptors, but do not directly measure pseudo-label quality or prove a particular error-feedback mechanism.

5 Conclusion

Selective Target Geometry uses saved source statistics as a starting point and gates target-side prototype and student updates with semantic and weak audio-visual evidence. Across the ten reported transfers, it improves MAE over the frozen model and five adaptation baselines; the router has the largest effect in the MOSI-to-SIMS ablation. The evidence is limited to one backbone, one task family, and offline adaptation. Related-corpus overlap and the generality of selective routing warrant further study.

AI Use Statement

AI tools assisted with literature discovery, research ideation, drafting, and language polishing. The authors selected the research questions and remain responsible for verifying experiments, claims, and citations before submission. AI tools were not used to prove mathematical claims.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Zirun Guo, Tao Jin, Wenlong Xu, Wang Lin, and Yangyang Wu. Bridging the Gap for Test-Time Multimodal Sentiment Analysis, February 2025. URL http://arxiv.org/abs/2412.07121.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1415fe9fea0fa1e45dddcff5682239a0-Paper.pdf.

Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information, 2023. URL https://arxiv.org/abs/2207.10792.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, 2021. URL https://arxiv.org/abs/2002.08546.

Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts. *International Journal of Computer Vision*, 133(1):31–64, January 2025. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-024-02181-w. URL http://arxiv.org/abs/2303.15361.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make Acoustic and Visual Cues Matter: CH-SIMS v2.0 Dataset and AV-Mixup Consistent Module, August 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting, 2022. URL https://arxiv.org/abs/2204.02610.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world, 2023. URL https://arxiv.org/abs/2302.12400.

Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation, 2021. URL https://arxiv.org/abs/2106.15326.

Kai Tang, Yixuan Tang, Tianyi Chen, Haokai Xu, Qiqi Luo, Jin Guang Zheng, Zhixin Zhang, Gang Chen, and Haobo Wang. Group-aware multiscale ensemble learning for test-time multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25831–25839, 2026. doi: 10.1609/aaai.v40i30.39782. URL https://doi.org/10.1609/aaai.v40i30.39782.

Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation, 2021. URL https://arxiv.org/abs/2103.14357.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences, 2019. URL https://arxiv.org/abs/1906.00295.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-time Adaptation by Entropy Minimization, March 2021. URL http://arxiv.org/abs/2006.10726.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/1905.12681.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Representation Learning*, volume 2024, pp. 55036–55055, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f25d75fc760aec0a6174f9f5d9da59b8-Paper-Conference.pdf.

Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Exploiting the intrinsic neighborhood structure for source-free domain adaptation, 2021. URL https://arxiv.org/abs/2110.04202.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.343. URL https://aclanthology.org/2020.acl-main.343/.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016. doi: 10.1109/MIS.2016.94.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis, 2017. URL https://arxiv.org/abs/1707.07250.

Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning, 2022. URL https://arxiv.org/abs/2211.06612.

A Detailed Method

A.1 Problem Formulation

We consider *test-time domain adaptation* (TTDA), the offline form of test-time adaptation Liang et al. (2025), for multimodal sentiment analysis. Let $\mathcal{S}$ and $\mathcal{T}$ denote the source and target domains. During *source training* we have labeled source data $\{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where $x_i = (x_i^T, x_i^A, x_i^V)$ comprises text, audio, and visual modalities; a fusion model $F_\theta = G \circ (E_T, E_A, E_V)$ with regression head C is trained on $\mathcal{S}$ together with the source-side statistics retained for adaptation (below, a class-conditional GMM). During *adaptation* the source data is inaccessible: we are given only the pretrained model (F_s, C_s), the shipped statistics, and the unlabeled target set $\{x_j^t\}_{j=1}^{N_t}$, which we may traverse for multiple epochs before producing final predictions—the transductive TTDA protocol of Liang et al. We write $\hat{F}$ and $\hat{C}$ for the adapting *student* model and head, initialized as copies of F_s and C_s.

The key challenge in **cross-lingual** adaptation—exemplified by cross-lingual transfer such as MOSI→SIMS—is that the conditional distribution shifts significantly:

$$P_s(Z \mid Y) \neq P_t(Z \mid Y), \quad P_s(Y \mid Z) \neq P_t(Y \mid Z). \tag{10}$$

The same label manifests through substantially different multimodal patterns across domains, so forcing target features to align with source-like prototypes Tian et al. (2021) can induce negative transfer.

A.2 Overview of Selective Target Geometry

We propose **Selective Target Geometry (STG)**, which balances source-domain prior knowledge with target-domain structural adaptation. During pretraining, STG builds a *source virtual domain*—a class-conditional Gaussian mixture fitted to source fusion features using the source labels. During adaptation, this geometry serves as the initial reference domain $\mathcal{U}_0$ and is EMA-updated *toward the target* batches; when a persistent mismatch is detected, additional *target-native* sub-domains are created alongside it, avoiding both the rigidity of pure source-alignment and the instability of initializing purely from noisy pseudo-labels.

As illustrated in Figure 2, STG comprises four layers:

1. **Semantic Anchor Layer** (§ *Semantic Anchor Layer*): A frozen text teacher provides reliability-aware soft pseudo-labels.
2. **Source-Anchored Geometry Layer** (§ *Source-Anchored Target Geometry*): Maintains class-conditional Gaussian prototypes, EMA-updated from target fusion features.
3. **Handcrafted Evidence Layer** (§ *Handcrafted Evidence Branch*): Groups handcrafted features into independent evidence models that assess pseudo-label quality without entering the model forward pass.
4. **Dynamic Router & Update Controller** (§ *Dynamic Router and Update Controller*): Decides per batch whether to *reuse*, *create*, or *defer* a virtual sub-domain and controls parameter updates.

A.3 SEMANTIC ANCHOR LAYER

Among the three modalities, text is the most semantically stable across domains: lexical sentiment cues survive changes of recording conditions, speaker population and annotation protocol far better than prosodic or facial cues, which are the channels through which most cross-lingual shift is expressed Tsai et al. (2019). This is consistent with the unimodal results on MOSI and MOSEI, where text is by far the strongest single modality Zadeh et al. (2016); Bagher Zadeh et al. (2018), and with the observation that jointly trained multi-modal networks are dragged toward their noisier inputs, so the cleanest channel is the safest anchor Wang et al. (2020). What matters for an anchor is not in-domain accuracy but transfer stability: the full multimodal prediction is carried in part by the audio-visual channels whose statistics shift most, and is therefore the less stable coordinate system under cross-lingual shift, whereas the text pathway degrades least. We therefore build the anchor from the *text pathway* of the frozen source model. Let $F_s^T(x^T)$ denote F_s restricted to text (audio and visual inputs masked, x^T propagated through E_T, the fusion stack and the head); teacher predictions are obtained by Monte Carlo dropout over this restricted map,

$$\begin{aligned} p_i^T &= \frac{1}{R}\sum_{r=1}^{R} F_s^T\big(x_i^T;\,\text{dropout}^{(r)}\big), \\ \sigma_i^T &= \text{std}\left(\{F_s^{T,(r)}(x_i^T)\}_{r=1}^{R}\right), \end{aligned} \tag{11}$$

where R is the number of stochastic passes and σ_i^T is the resulting predictive uncertainty. Regression outputs are converted to soft class labels by distance-based binning:

$$q_i^T(k) = \frac{\exp(-\gamma\,\|p_i^T - c_k\|^2)}{\sum_{l=1}^{K}\exp(-\gamma\,\|p_i^T - c_l\|^2)}, \tag{12}$$

where $\{c_k\}_{k=1}^{K}$ are pre-defined class centers ($c \in \{-1, 0, +1\}$, $K = 3$) and γ controls the sharpness of the binning. The teacher confidence is $w_i^T = \max_k q_i^T(k)$.

The teacher stays *frozen*, so erroneous pseudo-labels cannot corrupt the anchor through self-reinforcement. It fixes the class coordinate system $\{c_k\}$ shared by the virtual domain and the evidence models, and supplies the weight w_i^T that gates EMA updates and pseudo-label fusion. Note that the source virtual domain is *not* fitted from teacher outputs: at pretraining the ground-truth source labels are available and are used directly (Eq. 14); the teacher is used only at adaptation time, where labels are absent. The anchor is thus deliberately weaker in-domain than the multimodal student, but far more stable under shift.

A.4 SOURCE-ANCHORED TARGET GEOMETRY

STG maintains a library $\mathcal{V} = \{\mathcal{U}_e\}_{e=0}^{N_v}$ of virtual sub-domains, each storing class-conditional Gaussian distributions with diagonal covariance:

$$\mathcal{U}_e = \left\{(\mu_{e,k}, \Sigma_{e,k}, \pi_{e,k})\right\}_{k=1}^{K}. \tag{13}$$

Source domain initialization. The GMM is fitted during source training, when labels are available, and shipped with the checkpoint, as in other source-estimation TTA methods Tian et al. (2021); Qiu et al. (2021); Liang et al. (2025). Unlike VDM-DA Tian et al. (2021), which derives prototypes from classifier weights alone, STG fits $\mathcal{U}_0$ to the fusion features of the entire source training set:

$$\begin{aligned} \mu_{0,k} &= \frac{\sum_{i\in\mathcal{S}}\mathbb{I}[y_i \mapsto k]\; h_i}{\sum_{i\in\mathcal{S}}\mathbb{I}[y_i \mapsto k] + \epsilon}, \\ \pi_{0,k} &= \frac{\sum_{i\in\mathcal{S}}\mathbb{I}[y_i \mapsto k]}{N_s}, \end{aligned} \tag{14}$$

where $h_i = G(E_T(x_i^T), E_A(x_i^A), E_V(x_i^V))$ and $\mathbb{I}[y_i \mapsto k]$ maps a continuous label to its nearest class center. Covariances use Ledoit–Wolf-style shrinkage toward a spherical prior, $\Sigma_{0,k} = (1 - \lambda)\hat{\Sigma}_k + \lambda\sigma_0^2 I$. $\mathcal{U}_0$ is saved alongside the pretrained checkpoint and loaded at adaptation time as a statistically grounded reference geometry.

Target-native domain creation. When the router detects a persistently incompatible distribution (§ *Dynamic Router and Update Controller*), a new domain $\mathcal{U}_{N_v+1}$ is created from the buffered target samples using the same weighted-statistic formulation, with teacher soft labels q_i^T in place of the unavailable ground truth.

For a target sample with feature h_i, the class probability under domain $\mathcal{U}_e$ uses a temperature-scaled Mahalanobis distance:

$$d_{e,k}(h_i) = (h_i - \mu_{e,k})^\top \Sigma_{e,k}^{-1}(h_i - \mu_{e,k}), \tag{15}$$

$$q_{i,e}^V(k) = \frac{\pi_{e,k} \exp(-d_{e,k}(h_i)/\kappa)}{\sum_{l=1}^K \pi_{e,l} \exp(-d_{e,l}(h_i)/\kappa)}, \tag{16}$$

where the temperature $\kappa = D$ scales with the feature dimensionality so that the softmax neither saturates nor collapses as D changes. Throughout, $e^\star$ denotes the sub-domain selected by the router (Eq. 28); the virtual-domain prediction used downstream is $q_i^V \equiv q_{i,e^\star}^V$, and the virtual support score

$$s_i^V = \exp\big(-\min_k d_{e^\star,k}(h_i)/\kappa\big) \tag{17}$$

measures how well h_i fits that geometry.

Prototypes are updated by reliability-weighted EMA to prevent catastrophic drift:

$$\hat{\mu}_{e,k}^B = \frac{\sum_{i\in B} a_i \bar{q}_i(k) h_i}{\sum_{i\in B} a_i \bar{q}_i(k) + \epsilon}, \qquad \hat{\pi}_{e,k}^B = \frac{\sum_{i\in B} a_i \bar{q}_i(k)}{\sum_{i\in B} a_i + \epsilon}, \tag{18}$$

$$\mu_{e,k} \leftarrow (1-\beta_k)\,\mu_{e,k} + \beta_k\,\hat{\mu}_{e,k}^B, \qquad \pi_{e,k} \leftarrow (1-\beta_k)\,\pi_{e,k} + \beta_k\,\hat{\pi}_{e,k}^B, \tag{19}$$

with adaptive momentum $\beta_k = \beta_0 \cdot \mathrm{clip}(n_k^{\text{eff}}/n_0, 0, 1)$, where $n_k^{\text{eff}} = \sum_{i\in B} a_i \bar{q}_i(k)$ and $\beta_0 \approx 0.01$, so that no single batch can move a prototype far. Diagonal covariances follow the same EMA rule applied to the weighted second moment about $\mu_{e,k}$ plus the same shrinkage, so they cannot collapse. The reliability a_i is defined in § *Multi-Evidence Pseudo-Label Calibration*.

A.5 Handcrafted Evidence Branch

Prior work has attempted to use handcrafted features as hard constraints (e.g., "when F0 is high, the label cannot be neutral"), but our empirical analysis indicates that audio-visual handcrafted features (F0, MFCC, jitter, AU intensity, gaze, etc.) are **not** cross-domain invariants—their correlation with sentiment varies substantially across datasets. We therefore reformulate them as *interpretable weak evidence signals* that **never enter the model's forward pass**, serving solely as an independent assessment layer for pseudo-label quality and conflict detection.

We partition the 95 handcrafted features into $M = 9$ semantically grouped evidence models (Table 3), each modeling the class-conditional distribution of its feature subset with a diagonal Gaussian:

$$P_{\phi_g}\big(r_i^{(g)} \mid Y = k\big) = \mathcal{N}\big(r_i^{(g)}; \mu_{g,k}, \mathrm{diag}(\sigma_{g,k}^2)\big). \tag{20}$$

Source features are unavailable at adaptation time, so each group is fitted on a short warm-up segment of the target stream with teacher soft labels as pseudo-targets and refreshed by EMA. The evidence layer is therefore target-native but shares the teacher's label noise; we contain this by treating evidence as a *moderator* rather than a source of direction: a group that fails to corroborate the student has its ω_g driven toward zero, so inherited bias loses influence instead of propagating.

Group g produces a per-class log-evidence score as the log-ratio of the class-conditional likelihood to the prior-weighted sum of alternatives:

$$E_{i,g,k} = \log \frac{\pi_k\, P(r_i^{(g)} \mid Y = k)}{\sum_{l\neq k} \pi_l\, P(r_i^{(g)} \mid Y = l)}, \tag{21}$$

Table 3: Handcrafted evidence groups. Each group independently models $P(r^{(g)} \mid Y = k)$ using a Gaussian with diagonal covariance.

Group	Features	Dim
PITCH	F0 mean, std, min, max, range	5
ENERGY	RMS energy mean, std, min, max	4
TEMPO	Speech rate, pause ratio	2
VOICE QUALITY	Jitter, shimmer, HNR	6
MFCC-SPECTRAL	MFCC 0–12 + spectral centroid, bandwidth, rolloff, flatness	34
UPPER FACE AUS	AU01,02,04,05,06,07 (brow, cheek, lid)	12
LOWER FACE AUS	AU09,10,12,14,15,17,20,23,25,26,28 (nose, lip, jaw)	22
HEAD POSE	Rotation Rx, Ry, Rz statistics	6
GAZE & EYE	Gaze angle, eye openness	4
Total	9 independent evidence models	95

so that $E_{i,g,k} > 0$ exactly when group g favours class k over all alternatives jointly. Let $\hat{k}_i = \arg\max_k \hat{q}_i(k)$ be the class predicted by the *student*, where $\hat{q}_i$ is obtained from the student's regression output by the same distance-based binning as Eq. 12, and define the agreement indicator

$$\mathrm{agr}_{i,g} = \mathbb{I}\big[\arg\max_k E_{i,g,k} = \hat{k}_i\big] \in \{0, 1\}. \tag{22}$$

Each group maintains a reliability weight $\omega_g \in [0, 1]$ updated by EMA from prediction–evidence agreement over the current batch B:

$$\omega_g \leftarrow \eta\, \omega_g + (1 - \eta)\, \mathbb{E}_{i \in B}\big[\mathrm{agr}_{i,g}\big]. \tag{23}$$

The aggregated evidence is $E_i^{\text{hand}}(k) = \sum_{g=1}^{M} \omega_g \cdot E_{i,g,k}$. A per-sample consensus score measures how strongly the weighted evidence supports the student's own prediction:

$$\begin{aligned} \mathrm{cons}_i &= \mathrm{sig}\left(\sum_{g=1}^{M} \omega_g \cdot \mathrm{sign}\big(E_{i,g,\hat{k}_i}\big)\right), \\ \mathrm{conflict}_i &= 1 - \mathrm{cons}_i, \end{aligned} \tag{24}$$

where $\mathrm{sig}(\cdot)$ is the logistic sigmoid. High $\mathrm{conflict}_i$ means that the independent evidence contradicts the model prediction. Averaging over a batch gives the batch-level consensus $\bar{c}(B) = \frac{1}{|B|} \sum_{i \in B} \mathrm{cons}_i$ used by the router.

A.6 MULTI-EVIDENCE PSEUDO-LABEL CALIBRATION

STG fuses three complementary sources—text teacher q_i^T, virtual domain q_i^V, and handcrafted evidence E_i^{hand}—through a reliability-weighted product of experts:

$$\log \bar{q}_i(k) \propto \alpha_i^T \log q_i^T(k) + \alpha_i^V \log q_i^V(k) + \alpha_i^H E_i^{\text{hand}}(k), \tag{25}$$

with $\alpha_i^T = w_i^T$, $\alpha_i^V = s_i^V$, $\alpha_i^H = \lambda_H$, normalized over k. The sources are complementary rather than statistically independent—q_i^T and q_i^V both descend from the source model and the evidence models are fitted with teacher labels—but they *fail differently* (non-lexical sentiment, target geometry departing from the source, non-transferable feature–label relations), which is what makes disagreement informative.

The per-sample reliability combines confidence, geometric fit, teacher–virtual agreement and evidence consensus:

$$a_i = w_i^T \cdot s_i^V \cdot \big(1 - \overline{\mathrm{JS}}(q_i^T \,\|\, q_i^V)\big) \cdot (1 - \mathrm{conflict}_i), \tag{26}$$

where $\overline{\mathrm{JS}} = \mathrm{JS}/\log 2 \in [0, 1]$ is the Jensen–Shannon divergence normalized by its maximum, so that every factor of Eq. 26 lies in $[0, 1]$ and a_i does.

Each sample is then assigned one of three states:

- **Accept** if $a_i > \tau_{\text{accept}}$ and $\arg\max_k \bar{q}_i(k) = \hat{k}_i$; the fused label is used with weight a_i.
- **Correct** if $a_i > \tau_{\text{accept}}$ but the fusion disagrees with the student, *and* the two non-teacher sources agree with each other ($\arg\max_k q_i^V(k) = \arg\max_k E_i^{\text{hand}}(k)$); the fused label replaces the student's and is used with weight a_i.
- **Reject** if $a_i < \tau_{\text{reject}}$ or $\text{conflict}_i > 1 - \tau_{\text{cons}}$; the sample contributes to neither the loss nor the EMA update.

Accepted and corrected samples produce the regression target $\bar{y}_i = \sum_k c_k \, \bar{q}_i(k)$ used by $\mathcal{L}_{\text{pseudo}}$ in the supplementary material.

A.7 Dynamic Router and Update Controller

Given a batch B with fusion features $\{h_i\}$ and teacher soft labels $\{q_i^T\}$, the router scores the compatibility of B with each sub-domain:

$$C_e(B) = -\frac{1}{|B|} \sum_{i \in B} \sum_{k=1}^{K} q_i^T(k) \, \frac{d_{e,k}(h_i)}{\kappa} + \lambda_{\text{sep}} \cdot \min_{j_1 \neq j_2} \|\mu_{e,j_1} - \mu_{e,j_2}\|_2. \tag{27}$$

The first term is the negated teacher-weighted mean Mahalanobis distance, high when the batch sits close to the prototypes of $\mathcal{U}_e$ in the teacher's class coordinates. The second is a separation bonus over the closest prototype pair in $\mathcal{U}_e$; it is a property of the sub-domain, and stops the router preferring a collapsed domain that fits everything by being undiscriminative.

Let $\bar{C}(B) \equiv \max_e C_e(B)$ be the best compatibility in the batch and $\bar{c}(B)$ the batch-level evidence consensus of Eq. 24. The routing decision combines the two:

$$r(B) = \begin{cases} \text{DEFER}, & \bar{c}(B) < \tau_{\text{cons}}, \\ \text{REUSE}(e^\star), & C_{e^\star} > \tau_{\text{reuse}} \ \wedge \ \Delta C > \delta, \\ \text{CREATE}, & \bar{C}(B) < \tau_{\text{create}} \text{ for } L \text{ windows}, \\ \text{DEFER}, & \text{otherwise}, \end{cases} \tag{28}$$

where $e^\star$ and $e^{(2)}$ are the best- and second-scoring sub-domains, $\Delta C = C_{e^\star} - C_{e^{(2)}}$ their margin, and "for L windows" means the condition must hold for L consecutive batches. DEFER is thus reached by two routes—an explicit low-consensus gate and the residual branch—and in both cases the batch is buffered rather than used. CREATE requires low compatibility to persist for L consecutive windows, so transient fluctuations cannot spawn spurious domains, and REUSE requires a margin δ so that a batch is only absorbed by a sub-domain that is unambiguously the best match.

To prevent unbounded proliferation, STG enforces a *cooldown* of T_{cool} steps between two creations, a capacity bound $|\mathcal{V}| \leq N_{\max}$, periodic *merging* of highly similar domains ($\cos\text{-sim}(\mu_{e_1}, \mu_{e_2}) > \tau_{\text{merge}}$) every T_{merge} epochs, and *pruning* of domains unused for T_{idle} steps.

Unlike TTA methods that uniformly update normalization layers (e.g., TENT Wang et al. (2021)), STG modulates updates per routing outcome. REUSE applies the EMA of Eq. 19 to the selected sub-domain and steps the student's normalization layers at a reduced learning rate. CREATE fits a new sub-domain from the pending buffer at the full adaptation rate while existing sub-domains are held fixed, and skips the student update, since the batch was just judged incompatible with every existing domain. DEFER performs no update.

A.8 Training Objectives

The student is optimized with a composite loss of four terms: pseudo-label regression against accepted and corrected teacher outputs, virtual-prototype alignment (cross-entropy between the student's binned prediction and the stop-gradient virtual-domain posterior), compactness regularization (pulling high-reliability samples toward their assigned prototypes), and dropout-based consistency regularization. Full definitions are given in the supplementary material. The total loss is:

$$\mathcal{L}_{\text{TTA}} = \mathcal{L}_{\text{pseudo}} + \lambda_1 \mathcal{L}_{\text{proto}} + \lambda_2 \mathcal{L}_{\text{compact}} + \lambda_3 \mathcal{L}_{\text{cons}}. \tag{29}$$

The three coefficients $\lambda_1, \lambda_2, \lambda_3$ are the only tunable weights in the objective and are all small (λ_1=0.01, λ_2=0.005, λ_3=0.1), so that the pseudo-label regression term dominates and the structural terms act as regularizers. Every remaining hyperparameter—thresholds, momenta, window lengths and domain-management constants—is fixed to the same value across all ten scenarios and is listed in the supplementary material. The complete STG procedure is also given there.

B Training Objectives (Full Definitions)

The four loss terms constituting $\mathcal{L}_{\text{TTA}}$ are defined over a batch B of target samples. Only *Accepted* and *Corrected* samples contribute; let $a_i^{\text{acc}} = a_i \cdot \mathbb{I}[\text{state}_i \in \{\text{Accept}, \text{Correct}\}]$.

Pseudo-Label Regression Loss. The student's regression output is fitted to the calibrated pseudo-label:

$$\mathcal{L}_{\text{pseudo}} = \frac{\sum_{i \in B} a_i^{\text{acc}} \, \ell_1\big(\hat{C}(\hat{h}_i),\ \bar{y}_i\big)}{\sum_{i \in B} a_i^{\text{acc}} + \epsilon}, \tag{30}$$

where $\hat{h}_i$ is the student's fusion feature and $\bar{y}_i$ is the fused target defined above.

Virtual Prototype Loss. Aligns the student's binned prediction with the virtual-domain posterior. The posterior is treated as a constant (stop-gradient), so the gradient flows only into the student:

$$\mathcal{L}_{\text{proto}} = -\frac{\sum_{i \in B} a_i \sum_{k=1}^{K} q^V_{i,e^\star}(k) \, \log \hat{q}_i(k)}{\sum_{i \in B} a_i + \epsilon}. \tag{31}$$

Compactness Loss. Pulls high-reliability samples toward their assigned class prototypes:

$$\mathcal{L}_{\text{compact}} = \frac{1}{D} \cdot \frac{\sum_{i \in B} a_i \sum_{k=1}^{K} \bar{q}_i(k) \, d_{e^\star,k}(\hat{h}_i)}{\sum_{i \in B} a_i + \epsilon}. \tag{32}$$

Consistency Regularization. Implicit augmentation via two stochastic (dropout) forward passes of the student on the same sample:

$$\mathcal{L}_{\text{cons}} = \frac{1}{|B|} \sum_{i \in B} \ell_2\big(\hat{C}(\hat{h}_i),\ \hat{C}(\hat{h}'_i)\big), \tag{33}$$

where $\hat{h}_i$ and $\hat{h}'_i$ are the two stochastic passes.

C Algorithm

D Full Experimental Material

Data splits. MOSI: 1,284/229/686; MOSEI: 16,326/1,871/4,659; CH-SIMS: 1,368/456/457; CH-SIMSv2: 2,722/647/1,034 train/validation/test. Target training examples are unlabeled during adaptation.

E Implementation Details

Backbone and pretraining. proj_dim $D = 64$; 2-layer unimodal transformers; 3-layer cross-modal fusion; 3-layer MLP regression head. Pretraining: MSE loss, Adam, $lr = 10^{-3}$, batch size 64, 50 epochs, learning rate halved after 8 non-improving validation epochs, early stopping after 5. Unimodal encoders are frozen. The source virtual domain is fitted on the source training set with shrinkage $\lambda = 0.1$ and saved with the checkpoint.

Algorithm 1 STG: Selective Target Geometry

Require: Pretrained source model (F_s, C_s) and source virtual domain $\mathcal{U}_0$; unlabeled target data $\mathcal{D}_t$; handcrafted feature extractors Φ_A, Φ_V; epochs T; hyperparameters $\lambda_{1:3}, \beta_0, n_0, \kappa, \gamma, \eta, R, \lambda_H, \lambda_{\text{sep}}$; thresholds $\tau_{\text{cons}}, \tau_{\text{reuse}}, \tau_{\text{create}}, \tau_{\text{accept}}, \tau_{\text{reject}}, \tau_{\text{merge}}$; margin δ; persistence window L; domain management $N_{\max}, T_{\text{cool}}, T_{\text{merge}}, T_{\text{idle}}$.

Ensure: Adapted student $(\hat{F}, \hat{C})$.

$(\hat{F}, \hat{C}) \leftarrow \text{copy}(F_s, C_s)$; freeze $\{p \in \hat{F} \mid \text{“norm”} \notin \text{name}(p)\}$ *(cf. TENT)*
$\mathcal{V} \leftarrow \{\mathcal{U}_0\}$; $\mathcal{M} \leftarrow \emptyset$ *(pending buffer)*
Fit evidence models $\{\phi_g\}_{g=1}^{M}$ on the first n_{warm} target batches using teacher soft labels q_i^T; initialize $\omega_g \leftarrow 1$
for epoch $t = 1, \ldots, T$ **do**
  for each target batch $B \subset \mathcal{D}_t$ **do**
    Teacher: p_i^T, σ_i^T by MC dropout over F_s^T; q_i^T, w_i^T by distance-based binning
    Student: fusion features $\hat{h}_i$, binned prediction $\hat{q}_i$, $\hat{k}_i$
    Virtual domain: $q_i^V \equiv q_{i,e^\star}^V$ and s_i^V
    Evidence: E_i^{hand}, cons_i, conflict_i, batch consensus $\bar{c}(B)$
    Fuse $\bar{q}_i$ (product of experts); reliability a_i; $\bar{y}_i$
    Assign Accept / Correct / Reject per sample
    Routing action $r(B) \in \{\text{REUSE}, \text{CREATE}, \text{DEFER}\}$
    if $r(B) = \text{REUSE}$ **then**
      EMA-update $\mu_{e^\star,k}, \Sigma_{e^\star,k}, \pi_{e^\star,k}$ with momentum β_k
      Update student normalization layers with $\mathcal{L}_{\text{TTA}}$ at reduced learning rate
    else if $r(B) = \text{CREATE}$ **and** cooldown elapsed **and** $|\mathcal{V}| < N_{\max}$ **then**
      Spawn $\mathcal{U}_{N_v+1}$ from $\mathcal{M}$; fit its prototypes at the full adaptation rate; hold existing sub-domains and the student normalization layers fixed
      $\mathcal{M} \leftarrow \emptyset$
    else
      $\mathcal{M} \leftarrow \mathcal{M} \cup B$; no update *(defer)*
    end if
    Update evidence reliabilities $\{\omega_g\}$ by EMA
  end for
  if $t \equiv 0 \pmod{T_{\text{merge}}}$ **then**
    Merge sub-domains with $\cos\text{-sim}(\mu_{e_1}, \mu_{e_2}) > \tau_{\text{merge}}$; prune sub-domains unused for T_{idle} steps
  end if
end for
return $(\hat{F}, \hat{C})$

Adaptation. Adam, $lr = 10^{-4}$, batch size 64, $T = 30$ epochs, gradient clipping 1.0. Prototype/prior/covariance EMA base momentum $\beta_0 = 0.01$ (equivalently an EMA momentum of 0.99). Loss coefficients $\lambda_1 = 0.01$, $\lambda_2 = 0.005$, $\lambda_3 = 0.1$; evidence weight $\lambda_H = 0.05$. Softmax temperature $\kappa = D = 64$. Evidence reliability EMA $\eta = 0.9$; $M = 9$ groups. Routing thresholds $\tau_{\text{cons}} = 0.3$, $\tau_{\text{reuse}} = 0.6$, $\tau_{\text{create}} = -1.5$. Domain management: $N_{\max} = 3$, $\tau_{\text{merge}} = 0.9$. All of the above are identical for every one of the ten scenarios.

Evidence groups. 9 groups as detailed in Table 3. Each group uses a diagonal Gaussian model fitted from warm-up target batches with teacher soft labels and refreshed online together with its reliability weight ω_g ($\eta = 0.9$).

Evaluation. All metrics are computed on the rescaled $[-1, +1]$ target scale. MAE is the mean absolute difference between predicted and ground-truth scores; Acc-2 and F1 are obtained by binarizing predictions and labels at zero; Acc@0.5 is the fraction of samples with $|\hat{y} - y| \leq 0.5$.

Compute. All experiments run on a single NVIDIA RTX 5090 (32GB). Pretraining takes 2–10 minutes per dataset. Adaptation takes 1–10 minutes per scenario (depending on target dataset size). Handcrafted feature extraction from raw data is performed once and cached.

Table 4: **Hyperparameter sensitivity.** Each coefficient or threshold is varied one at a time around its default ($1\times$ / τ_{cons}=0.3 / τ_{reuse}=0.6) on one cross-lingual (S1) and one same-language (S6) scenario; MAE ↓ and Acc-2 ↑ on the target test split.

Setting	S1 MAE	S1 Acc-2	S6 MAE	S6 Acc-2
λ_{proto} $0.5\times$	0.5685	53.02	0.4404	67.66
λ_{proto} $1\times$ (default)	0.5276	58.15	0.4149	70.21
λ_{proto} $2\times$	0.5290	56.79	0.4270	69.83
λ_{compact} $0.5\times$	0.5288	57.28	0.4167	70.95
λ_{compact} $1\times$ (default)	0.5276	58.15	0.4149	70.21
λ_{compact} $2\times$	0.5517	55.49	0.4544	68.13
λ_{cons} $0.5\times$	0.5396	57.86	0.4333	69.71
λ_{cons} $1\times$ (default)	0.5276	58.15	0.4149	70.21
λ_{cons} $2\times$	0.5312	58.21	0.4232	69.58
τ_{cons} 0.2	0.5434	56.25	0.4326	70.20
τ_{cons} 0.3 (default)	0.5276	58.15	0.4149	70.21
τ_{cons} 0.4	0.5570	55.36	0.4210	69.98
τ_{reuse} 0.4	0.6307	53.99	0.4395	67.24
τ_{reuse} 0.6 (default)	0.5276	58.15	0.4149	70.21
τ_{reuse} 0.8	0.5513	56.97	0.4386	68.57

Reproducibility. All reported numbers are averaged over multiple random seeds; the adaptation protocol, checkpoint selection and hyperparameters are identical across seeds and across methods.